REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302, and to the Office of Management and Budget, Paperwork Reduction Project (0704-0188), Washington, DC 20503.

1. AGENCY USE ONLY (Leave blank)	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
		FINAL REPORT 01 Jan 93 - 31 Dec 95

4. TITLE AND SUBTITLE

System Design and Architecture of Optoelectronic Smart Networks

5. FUNDING NUMBERS

61102F
2305/DS

6. AUTHOR(S)

Professor Kiamilev

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Department of Electrical Engineering
University of North Carolina at Charlotte
Charlotte, NC 28223

AFOSR-TR-96
0263

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)

AFOSR/NE
110 Duncan Avenue Suite B115
Bolling AFB DC 20332-0001

10. SPONSORING / MONITORING AGENCY REPORT NUMBER

F49620-93-1-0113

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION / AVAILABILITY STATEMENT

APPROVED FOR PUBLIC RELEASE: DISTRIBUTION UNLIMITED

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 words)*

This research program focused on the development of optoelectronic interconnection networks that combine communication and processing capabilities in network hardware to accelerate distributed computing applications.

19960614 074

14. SUBJECT TERMS

15. NUMBER OF PAGES

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
UNCLASSIFIED	UNCLASSIFIED	UNCLASSIFIED	

Final Report

Principal Investigator:	Dr. Fouad Kiamilev Assistant Professor of Electrical Engineering
Institution:	Department of Electrical Engineering University of North Carolina at Charlotte Charlotte, NC 28223

AFOSR Grant #:	F49620-93-1-0113
Program Title:	***System Design and Architecture of Optoelectronic Smart Networks***
Contract Period:	01 Jan. 93 through 31 Dec. 95 (36 Months)
Funding:	$48,000 for 3 years

1.1. Executive Summary

This research program focused on the development of optoelectronic interconnection networks that combine communication and processing capabilities in network hardware to accelerate distributed computing applications.

On the architecture front, we have designed an optoelectronic hardware module that can be used as a building block of smart networks with application-specific performance and cost requirements. A detailed technological comparison between the optoelectronic design and an equivalent advanced electronic MCM implementation was carried out to assess the advantages provided by the optoelectronic solution. The results of our work are incorporated into a prototype optoelectronic switch currently being built at Bell Laboratories, a division of Lucent Technologies.

On the hardware front, we have collaborated with Bell Laboratories to demonstrate a 2Kbit, 50Mpage/s, photonic first-in, first-out page buffer based on GaAs/AlGaAs multiple quantum well diodes flip-chip bonded to sub-micron CMOS circuits. This photonic chip provided a number of breakthroughs in the area of optical interconnects, including:

- *First implementation of a hybrid 850nm GaAs MQW/CMOS VLSI circuit with modulators bonded directly on top of active silicon circuits.*
- *First demonstration of a hybrid 850nm GaAs MQW/CMOS transimpedance transmitter/receiver circuit operating at 375 Mbis/sec with switching energy of ≈370 femto-joules.*
- *Design and implementation of a high-density 2Kbit photonic first-in, first-out page buffer circuit with optical input and output functionality.*
- *Measurement of ring oscillator circuits loaded with hybrid MQW devices operating at 2GHZ*

1.2. Personnel Supported

Dr. Fouad Kiamilev - Asst. Professor of Electrical Engineering at UNCC. Dr. Kiamilev is the principal investigator on this project.

Iyad Hadba - Mr. Hadba is an MS student in the Electrical Engineering Department at UNC Charlotte. He is partially supported by this contract. Majority of his funding comes from Rockwell International Digital Communications Division in the form of an internship. Mr. Hadba is finishing thesis "VHDL Design for a Microcontroller," that is related to the controller for the interconnection networks being considered under this program.

Ravi Narala - Mr. Narala was an MS student in the Electrical Engineering Department at UNC Charlotte. He was only partially supported by this contract. Majority of his funding came from MX-COM Inc. in the form of an internship. Mr. Narala's M.S. thesis "Design of an ATM segmentation chip" was an important part of this project in helping us to understand the hardware complexity associated with network interfaces. Mr. Narala is currently working as a chip designer for MX-COM Inc. in North Carolina.

Gordon Aplin - Ph.D. student in the department of Electrical Engineering at UNCC. Mr. Aplin is a United States citizen. Mr. Aplin was funded from a related AASERT (F49620-93-1-0472)

Dr. Ashok Krishnamoorthy - Member of Technical Staff, AT&T Bell Laboratories. We are collaborating with Dr. Krishnamoorthy on both theoretical and hardware portions of this program. Dr. Krishnamoorthy is not funded by this program.

Dr. Sadik Esener - Professor of Electrical and Computer Engineering at UCSD. We are collaborating with Dr. Esener on the theoretical portion of this program. Dr. Esener is not funded by this program.

1.3. Publications

- F. Kiamilev, E. Stevens, A.V. Krishnamoorthy and S.C. Esener, "Modular Architectures for Optoelectronic Multistage Interconnection Networks," International Journal of Optoelectronics, volume 9, number 6, pages 457-470, September 1994.
- J. Fan, B. Catanzaro, F. Kiamilev, S. Esener, and S.H. Lee, The Architecture of an Integrated Computer Aided Design System for Optoelectronics, Optical Engineering, Vol. 33 No. 5:1571-1580 (1994).
- A.V. Krishnamoorthy and F.E. Kiamilev, "Fanout, Replication and Buffer Sizing for a Class of Self-Routing Packet-Switched Multistage Photonic Switch Fabrics," in Proceedings of OSA Topical Meeting on Photonic Switching, Salt Lake City, page 87, March 1995.
- A.V. Krishnamoorthy, A.L. Lentine, K.W. Goosen, J.A. Walker, T.K. Woodward, J.E. Ford, G.F. Aplin, L.A. D'Asaro, S.P. Hui, B. Tseng, R. Leidenguth, D. Kossives, L.M.F. Chirovsky, and D.A.B. Miller, "3-D Integration of MQW Modulators over Active Sub-micron CMOS Circuits: 375Mb/s Transimpedance Receiver-Transmitter Circuit," Accepted for publication in IEEE Photonics Technology Letters, June 1995.
- A.V. Krishnamoorthy, J.E. Ford, K.W. Goosen, J.A. Walker, A.L. Lentine, S.P. Hui, B. Tseng, L.M.F. Chirovsky, R. Leibenguth, D. Kossives, D. Dahringer, L.A. D'Asaro, F.E. Kiamilev, G.F. Aplin, R.G. Rozier and D.A.B. Miller, "Photonic Page Buffer Based on GaAs MQW Modulators Bonded Directly Over Active Silicon CMOS Circuits," Submitted to Applied Optics, Special Issue on Optical Memory, August 1995.
- A.V. Krishnamoorthy, T.K. Woodward, R.A. Novotny, K.W. Goosen, J.A. Walker, A.L. Lentine, L.A. D'Asaro, S.P. Hui, B. Tseng, R. Leibenguth, D. Kossives, D. Dahringer, L.M.F.

Chirovsky, G.F. Aplin, R.G. Rozier, F.E. Kiamilev, and D.A.B. Miller, "Ring Oscillators with Optical and Electrical Readout based on Hybrid GaAs MQW Modulators Bonded to 0.8micron Silicon VLSI Circuits," Electronic Letters, Vol. 31, No. 22, pages 1917-1918, October 1995.

- A.V. Krishnamoorthy, J.E. Ford, K.W. Goosen, J.A. Walker, A.L. Lentine, T.K. Woodward, L.A. D'Asaro, S.P. Hui, B. Tseng, R. Leibenguth, D. Kossives, D. Dahringer, L.M.F. Chirovsky, F.E. Kiamilev, G.F. Aplin, R.G. Rozier and D.A.B. Miller, "3-D Integration of MQW SEED Detectors and Modulators over Active Sub-micron CMOS circuits: Application to a 2kbit Parallel Photonic Page Buffer," in Proc. LEOS Annual Meeting, Optical Interconnects and Processing Systems, San Francisco, October 1995.
- A.V. Krishnamoorthy, J.E. Ford, K.W. Goosen, J.A. Walker, A.L. Lentine, L.A. D'Asaro, S.P. Hui, B. Tseng, R. Leibenguth, D. Kossives, D. Dahringer, L.M.F. Chirovsky, F.E. Kiamilev, G.F. Aplin, R.G. Rozier and D.A.B. Miller, "Implementation of a Photonic Page Buffer Based on GaAs MQW Modulators Bonded Directly over Active Silicon VLSI Circuits," in Proc. OSA Topical Meeting on Optical Computing, Salt Lake City, Paper OTuD2 (postdeadline), March 1995.
- J.E. Ford, A.V. Krishnamoorthy, K.W. Goosen, J.A. Walker, D.A.B. Miller, R. Morrison, A.L. Lentine, S.P. Hui, B. Tseng, L.M.F. Chirovsky, R. Leibenguth, D. Kossives, D. Dahringer, L.A. D'Asaro, G.F. Aplin, R.G. Rozier and F.E. Kiamilev, "Optical Test of an Photonic FIFO Page Buffer Memory," to appear in Proc. OE/LASE'96 Conference on Optoelectronic Interconnects and Modules, San Jose, CA 1996.
- Implementation of a Photonic Page Buffer Based on GaAs MQW Modulators Bonded Directly over Active Silicon VLSI Circuits," in Proc. OSA Topical Meeting on Optical Computing, Salt Lake City, Paper OTuD2 (postdeadline), March 1995.
- J. Morris, W. Heyward, H. Yang, F. Kiamilev, Y. Raja, and M.R. Feldman, "Experimental demonstration of free-space optical interconnects for multichip modules," in OSA Annual Meeting (1994)
- A.V. Krishnamoorthy and F.E. Kiamilev, "Architecture and Performance of the Stretch Network," in Proc. 1994 SPIE Annual Meeting, San Diego, CA (1994)
- F.E. Kiamilev and A.V. Krishnamoorthy, "Smart Pixel Designs for Image Processing," in Proc. 1994 SPIE Annual Meeting, San Diego, CA (1994)
- F.E. Kiamilev, J.E. Morris, J. Childers, R. Sharma, V. Badoni, M.R. Feldman, Optically interconnected MCMs for ATM switches, in Optoelectronic Interconnects, Ray T. Chen, Editor, Proc. SPIE 1849, pp. 160-171 (1993)
- F.E. Kiamilev, A.V. Krishnamoorthy, and S.C. Esener, Modular Architecture for Smart Pixel Switching Networks, in Optoelectronic Interconnects, Ray T. Chen, Editor, Proc. SPIE 1849, pp. 129-140 (1993)
- R. Sharma, F. Kiamilev, J. Morris, J. Childers, V. Badoni, M. Feldman, Multiplexed optical interconnects for multichip modules, OSA Annual Meeting Technical Digest, 1993 (Optical Society of America, Washington D.C. 1993), Vol. 16, pg. 190.
- J. Morris, W. Heyward, M. Nakkar, F. Kiamilev, Y. Raja, and M. Feldman, Multichip module utilizing free-space optical interconnects, OSA Annual Meeting Technical Digest, 1993 (Optical Society of America, Washington D.C. 1993), Vol. 16, pg. 43.

- F.E. Kiamilev, C. Graham, and P. Abiprojo, Optoelectronic multislotted interconnection networks, OSA Annual Meeting Technical Digest, 1993 (Optical Society of America, Washington D.C. 1993), Vol. 16, pg. 77.
- F.E. Kiamilev, R. Sharma, J. Childers, J. Morris, H. Yang, and M. Feldman, Optically interconnected MCMs for asynchronous-transfer-mode networks, OSA Annual Meeting Technical Digest, 1993 (Optical Society of America, Washington D.C. 1993), Vol. 16, pg. 10.
- F.E. Kiamilev and A.V. Krishnamoorthy, Smart Pixel Designs for Image Processing, in Proc. Optoelectronic Enhancements to Digital Computing Technology, SPIE Ann. Meet., San Diego, 1994.

1.4. New Discoveries, Inventions, or Patent Disclosures

- A.V. Krishnamoorthy and F.E. Kiamilev, *"Packet Switched Extended Generalized-Shuffle Self-Routing Multistage Interconnection Networks,"* Patent Application filed with the US Patent and Trademark Office, Serial No. 08/184,551. (1994).
- K.W. Goosen, F.E. Kiamilev, A.V. Krishnamoorthy, and D.A.B. Miller, *"Virtual Smart Pixel Design Methodology,"* Bell Laboratories Patent Submission. Disclosure Date: August 1994.
- J.E. Ford, F.E. Kiamilev, A.V. Krishamoorthy, "An Optoelectronic Switch for Multiprocessor Networking," Bell Laboratories Patent Submission. Disclosure Date: March 1996.

2. Smart Network Architecture

This section describes three multistage interconnection network (MIN) architectures that we have developed for use with optoelectronic technology. These architectures target different applications and therefore have distinct cost, performance and functionality characteristics. All the architectures can be built using a common hardware module. This module is a perfect shuffle MIN that utilizes simple processing elements (less than 200 logic gates per PE). Previously, we have shown that optoelectronic (e.g. smart-pixel) implementation of such modules is highly efficient, while electronic implementations suffer in scalability due to the large number of wire crossovers required in the perfect shuffle interconnection and their length[1,2]. Our approach allows a single optoelectronic packaging scheme to be applied in several applications, thus lowering the hardware development cost and increasing the number of potential users.

2.1. Introduction

In the past decade much interest has been generated in the use of self-routing multistage interconnection networks (MINs) for high-performance packet-switched interconnection networks for telecommunications in the form of asynchronous transfer mode (ATM) switches and internal networks for massively parallel computers[3,4]. The basic appeal of MINs lies in their implicit simplicity and their scalability to a large number of ports. Unfortunately, the scalability potential of electronic implementations of these networks is often overshadowed by physical packaging constraints in the form of limited chip pin-outs, connector limitations on PCB's, backplane interconnection density, and for high-speed systems, signal integrity and latency characteristics[2,5].

Previously, a number of researchers have proposed that MINs can be efficiently implemented using 2-D optoelectronic processor arrays interconnected with free-space optical links[6,7,8,9,10,11,12].

Moreover, several research groups are building system prototypes of such optoelectronic MINs[13]. For example, we have designed an optoelectronic perfect shuffle MIN and shown that this design outperforms both chip-level and multichip module (MCM) level electronic implementations[2,14].

In this program, we have used our previous hardware design as the building block (or module) for three interconnection network architectures with distinct application, performance and cost requirements. The first architecture, called the tandem banyan network, provides one-to-one communication and is well suited for uniform traffic (e.g. all output ports are equally likely to be used). While this architecture achieves low cost and low latency, it has very poor performance in the presence of "hot-spot" traffic (where some output ports are more likely to be used) and it does not support one-to-many (or broadcast) communication. The second architecture, called the smart network, solves these problems at the expense of higher hardware cost and higher latency. The third architecture, called the hierarchical network, reduces the smart network latency at the expense of higher hardware cost.

2.2. *Application Requirements*

This section describes the application targeted by the architectures presented in this paper. Nominally, this application is ATM which is emerging as a standard networking scheme for high-speed computer interconnection both at LAN and WAN levels[15]. In the near future, ATM networks will carry traditional computer data traffic, such as data files and email messages, as well as video, voice and data traffic associated with distributed computation performed by a group of processors attached to the network.

The experimental switch activities of telecommunications vendors is currently focused (for the most part) on electronic ATM switch systems providing small numbers of ports (typically 16-64) operating at 155 and 622 Mbps each. Some Japanese telecommunication vendors have demonstrated switches operating at 1.8 Gbps and 2.5Gbps implemented with GaAs ICs and advanced MCM packaging technologies[16,17,18]. All these systems are typically based on the crossbar, shared memory, or shared medium (e.g. bus and ring) switch architectures. While these architectures are adequate for today's networking applications, scaling them to meet future switching demands will present a formidable challenge. The challenge is to efficiently implement switches with large number of physical ports (1K-10K ports) operating at gigabit data rates (1-10 Gbps/port) and having 1 to 10 terabit/second aggregate bandwidth capacities.

There are substantiated engineering tradeoffs to take into consideration when deciding on a switch architecture that has to scale to over 1000 physical ports and operate at gigabit port data rates. Physical packaging issues become very important. Technologies, architectures and systems which have worked well for a 64 port switch operating at 155Mbps, are often impractical for a 1000 port switch operating at 1Gbps. For example, both the interconnect and circuit complexity of a crossbar switch architecture grows as $O(N^2)$, making it impractical for network sizes of 1000 and above. Likewise, both shared memory and shared medium architectures suffer from a performance degradation as the number of channels is increased.

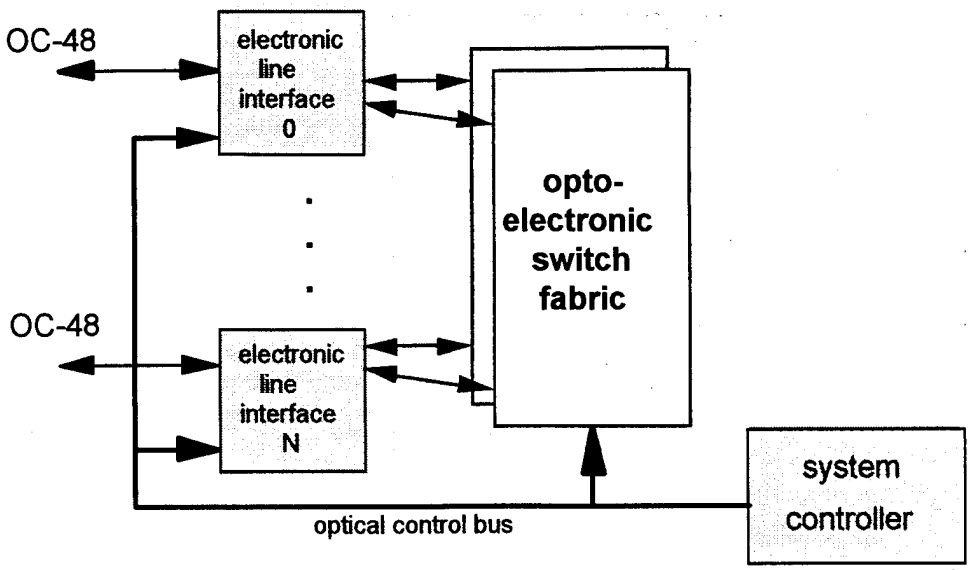


Figure 1. A switching network consists of the switch fabric, line interfaces and controller.

Figure 1 shows a typical switching network architecture where the network nodes (e.g. processors, memories, or specialized devices) are attached to the interconnection network (or switch fabric) via line interfaces. On the input side, line interfaces segment incoming data messages into fixed size cells (or packets) for transmission through the switch fabric. The self-routing switch fabric routes these cells between input and output ports. On the output side, line interfaces reassemble the original data message. In addition to cell segmentation and reassembly, line interfaces also incorporate I/O interface, cell buffering, and network protocol functions. Finally, the system controller is used for higher level functions such as network management and testing.

Although line interfaces are an important part of a complete network design, the architectures described in this paper implement the switch fabric portion of the switching network. Design issues for switch fabrics include types of provided communication services, cell blocking (or cell loss rate), guaranteed cell delivery, latency and cell priority. Typically, switch fabrics are engineered to meet application-specific performance and cost requirements. For example, in this paper we describe three switch fabric architectures that target distinct performance and cost requirements.

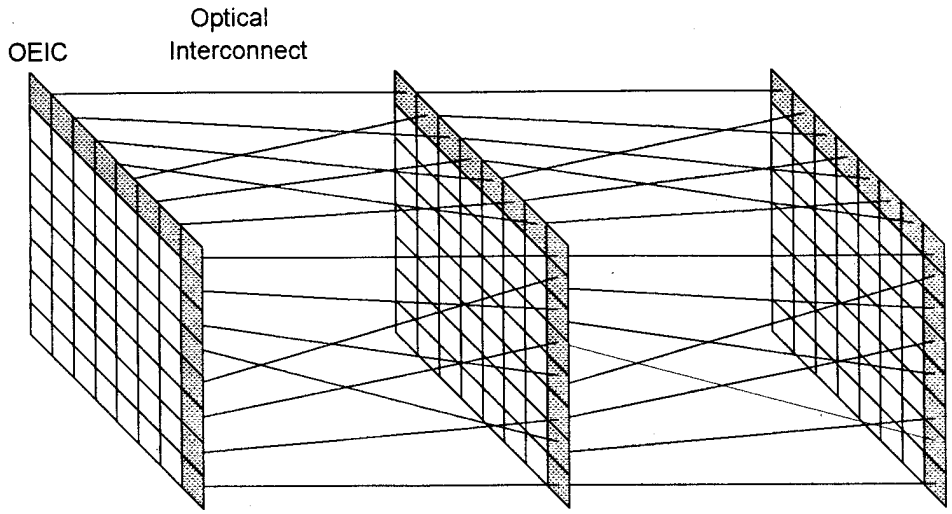
The two types of cell blocking that can occur in ATM switch fabrics are internal link blocking and output port blocking. Internal link blocking occurs for switch fabrics that cannot support all possible interconnections. In this situation, it is possible that two cells will simultaneously compete for the same link and one of the cells has to be discarded or buffered for later transmission. Output port blocking is unavoidable in self-routing switch fabrics because several input ports can simultaneously send a cell to the same output port. Typically, networks are engineered to allow small amounts of blocking (e.g. $\leq 10^{-9}$) for a given distribution of incoming traffic (e.g. uniform, community of interest, bursty, etc.).

Guaranteed cell delivery is critical for multimedia applications that require a sustained bandwidth to be maintained between the network devices at any time during the connection. This is closely related to cell priority which allows cells with higher priority to have precedence over cells with lower priority (e.g. high priority traffic is delivered first) when blocking occurs. Finally, latency is important when the switch fabric is used for distributed computing, in which case lower communication latency yields more efficient parallel processing.

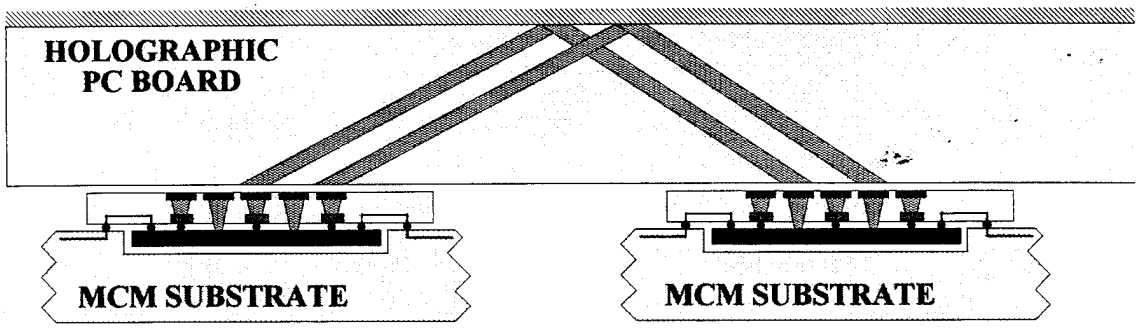
Cell traffic through the switch fabric can be divided in two categories: communication traffic and synchronization traffic. Communication traffic transfers cells between input and output ports. Typical communication traffic consists of one-to-one and one-to-many (e.g. multicast or broadcast) cell traffic. In one-to-one traffic, cells are send from a source port to a single destination port. In multicast traffic, a source port simultaneously sends the same cell to many destination ports. Synchronization traffic occurs in distributed applications[19] where a software program is partitioned into a set of cooperating processes that run concurrently on different processors and communicate using message-passing over the interconnection network. Unless properly handled, synchronization traffic can lead to "hot-spot" network congestion[20] as described in the following example.

To illustrate synchronization traffic, consider a parallel implementation of a loop with M iterations, followed by a sequential code portion. We can have M processors executing the M iterations of the loop in parallel, but the sequential portion of the code has to wait until all M processors are finished. In a shared memory computer, this type of synchronization is implemented by having each processor increment a shared memory variable. The processor containing the serial code checks the variable to decide when it can execute. The problem arises when all the processors finish and send M messages to increment the same shared variable. Since the interconnection network has only one output port to the memory containing the shared variable, the updates must be done serially, creating a performance bottleneck. This phenomenon is called the synchronization bottleneck[21] (or MSYPS limit).

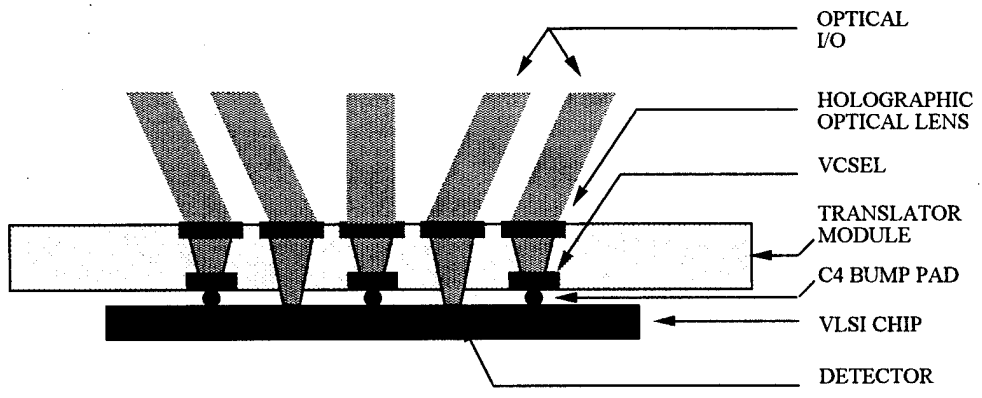
One approach to eliminating the synchronization bottleneck is not to parallelize the code that requires extensive use of synchronization operations. This approach cannot be used in distributed computing, because synchronization operations are inherent in distributed systems and are used for parallel resource scheduling and allocation[22]. Thus a method of efficiently performing synchronization has to be implemented in the network hardware to allow high-performance distributed computing.



(A)



(B)



(C)

Figure 2. Optoelectronic hardware module used to construct the MIN architectures presented in this paper. Figure A shows the schematic diagram of the design, while figures B and C show the physical implementation using optically interconnected multichip modules (reference 2).

2.3. Optoelectronic Hardware Module

Figure 2 shows the optoelectronic hardware module that is used as a building block of the architectures described in this paper. It consists of optoelectronic processor arrays connected in tandem using free-space optical links. The free-space links always use a 2-D shuffle exchange topology. Each processor array contains identical and simple processing elements (less than 200 logic gates/PE) uniformly spaced on a 2-D grid and having optical input and output ports. In our design approach, the logic function performed by the processing elements is architecture dependent, while the optical interconnection remains fixed. For example, although the architectures described in this paper use 8 distinct processor array types, they all rely on the same 2-D optical shuffle exchange to interconnect the arrays. This approach allows the same optoelectronic packaging scheme to be used to build the entire interconnection network thereby reducing fabrication and design costs.

The detailed hardware design of this module was previously described in references 2 and 14 and will not be repeated here. Cascading $\log_K N$ module stages produces a network that is functionally equivalent to an N channel perfect shuffle MIN, where K is a design parameter. The perfect shuffle MIN, shown in figure 3, uses $\log_2 N$ stages of switching elements. Each stage contains N/2 switching elements with 2 input and 2 output ports (see figure 4). Cells enter the switching elements in a particular stage, bit and frame aligned. Each switching element receives two incoming cells, examines the information contained in their cell headers, and routes them to the appropriate output port.

2.4. The Tandem Banyan Network

The tandem banyan MIN architecture was originally developed for electronic implementation (see figure 5). The basic idea behind this architecture is to repeat cell routing through a banyan network and after each routing attempt, remove cells that have been successfully routed. The tandem banyan network can be built using our optoelectronic hardware module because it is based on a topology that is equivalent to the perfect shuffle. The tandem banyan MIN provides one-to-one communication and cell priority services (e.g. higher priority cells have lower latency). It is well suited for local area computer networks, because of relatively low latency and low cost. A detailed description of the tandem banyan design and performance can be found in reference [23].

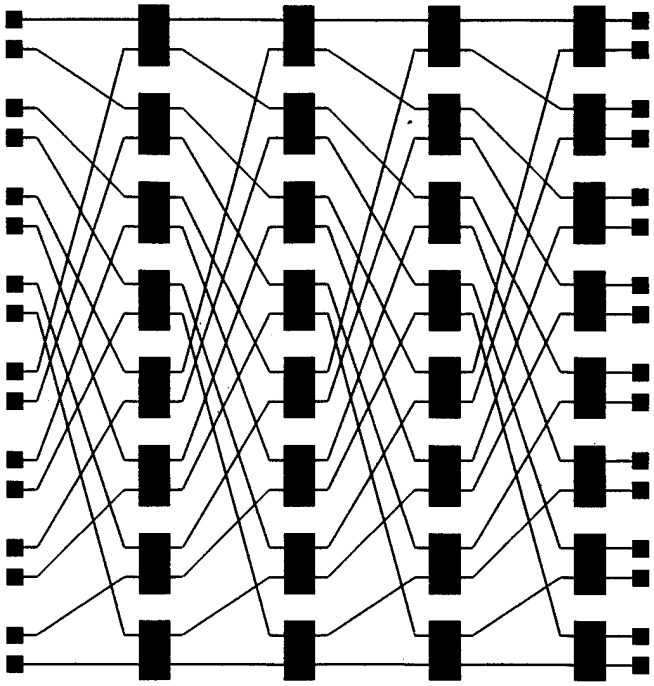


Figure 3. The perfect shuffle MIN architecture uses $\log_2 N$ stages of switching elements. Each stage contains N/2 switching elements with 2 input and 2 output ports. This figure is for N=16.

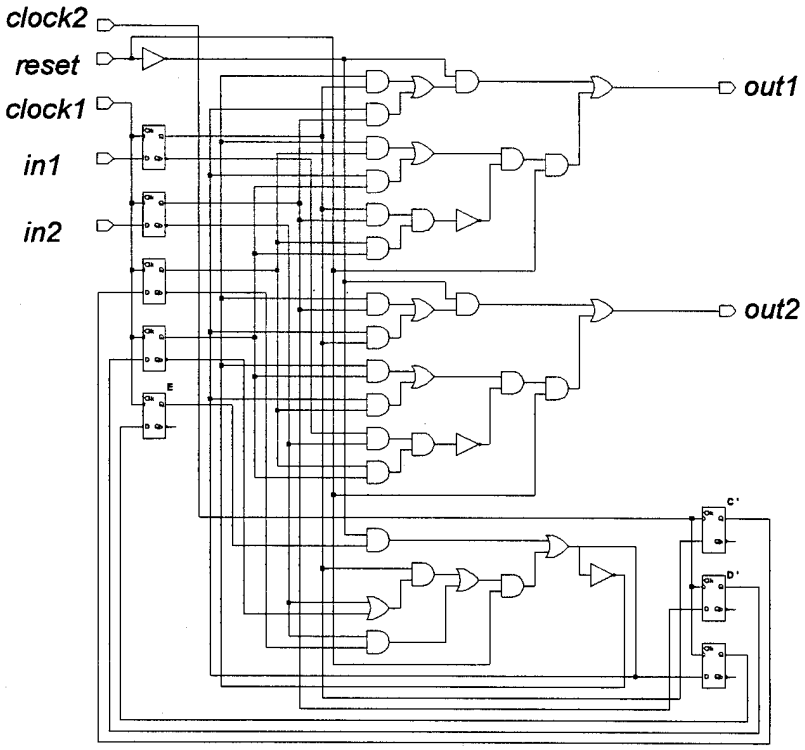


Figure 4. Gate level design of the 2x2 switching element for a typical perfect shuffle routing MIN. This design requires less than 100 logic gates.

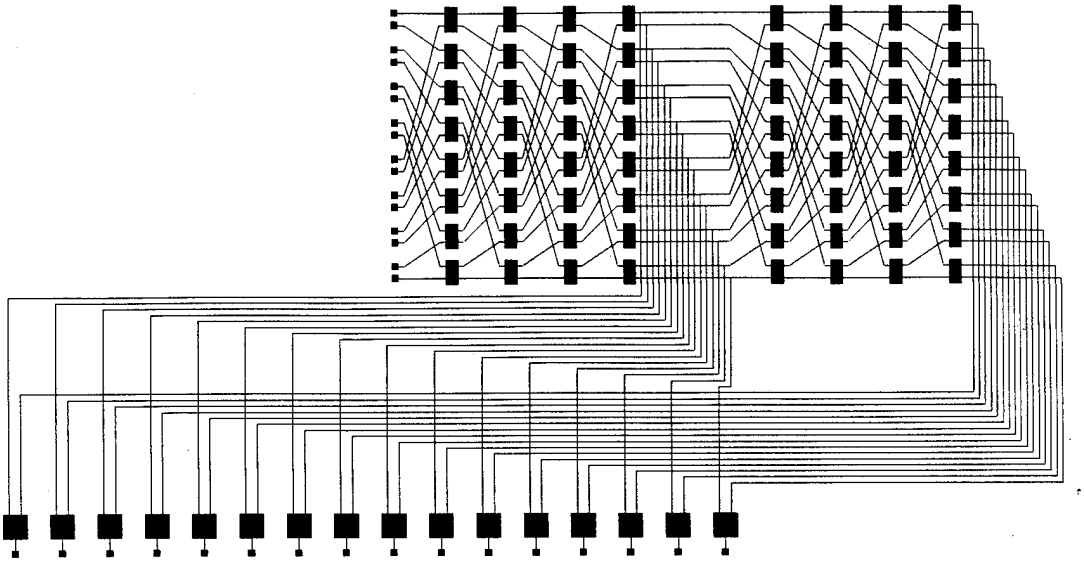


Figure 5. The tandem banyan MIN architecture cascades R perfect shuffle MINs. This figure shows the tandem banyan for R=2 and N=16.

Figure 6 shows the cell loss rate (defined as the probability of a cell being misrouted due to internal link blocking) for a 1024 port tandem banyan network as a function of the number of banyans in tandem (R). It can be seen that the cell loss rate can be made arbitrarily small by increasing R. For example, for R=8, the cell loss rate is near 10^{-5} and the number of stages is 80. Assuming that each banyan stage has a latency of 3 clock cycles (e.g. 1 cycle for the activity bit, 1 cycle for the priority bit, and 1 cycle for the routing bit), then the worst case latency of an R=8 tandem banyan network is 240 clock cycles (e.g. $8_{banyans} \cdot 10^{stages/banyan} \cdot 3^{cycles/stage}$). On the other hand the best case latency is 30 clock cycles (e.g. $1_{banyan} \cdot 10^{stages/banyan} \cdot 3^{cycles/stage}$). The average latency is 90 clock cycles (or 3 banyans in tandem) as determined by computer simulation.

A major shortcoming of the tandem banyan network is its inability to handle "hot-spot" traffic. Figure 7 shows the curves for cell loss rate of a 1024 tandem banyan network where 5% and 10% of all incoming cells are directed to a single output port while the remaining cells are uniformly distributed. It can be seen that the cell loss rate with "hot spot" traffic is much higher than the cell loss rate with uniform traffic (superimposed on the same plot). In fact, the "hot-spot" cell loss rate saturates near 10^{-1} even as R is increased to 10. This leads us into the next section, which enhances the tandem banyan network to resolve "hot-spot" traffic that arises due to synchronization operations.

2.5. The Smart Network

This section describes a new MIN architecture, called the smart network. The smart network architecture retains all the capabilities of the tandem banyan MIN while improving performance under "hot spot" traffic conditions and allowing broadcast communication. These additional capabilities are achieved at the expense of increased hardware costs and higher latency. The smart network is constructed by repeatedly cascading the basic optoelectronic hardware module described in section 3. Unlike the tandem banyan network, which uses a single processor array chip design, the smart network requires 7 different processor array designs.

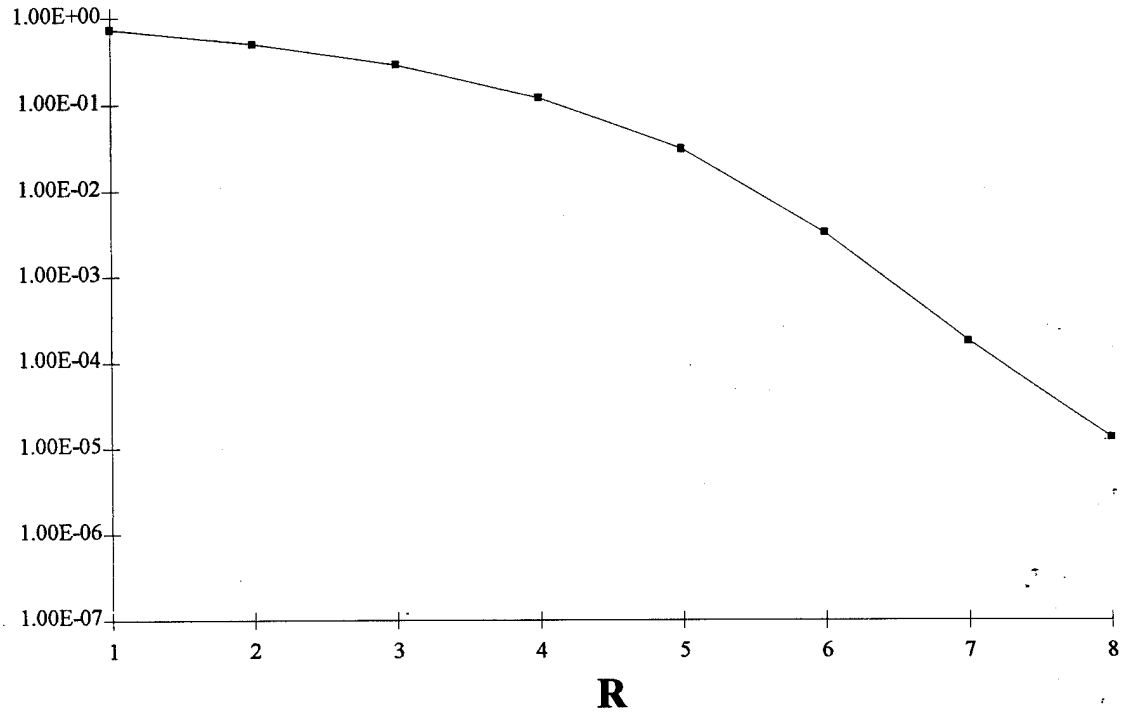


Figure 6. Cell loss rate under uniform traffic for N=1024 port tandem banyan network. The vertical axis is the cell loss rate, while the horizontal axis is R, the number of banyans in tandem. Input link load is 100%.

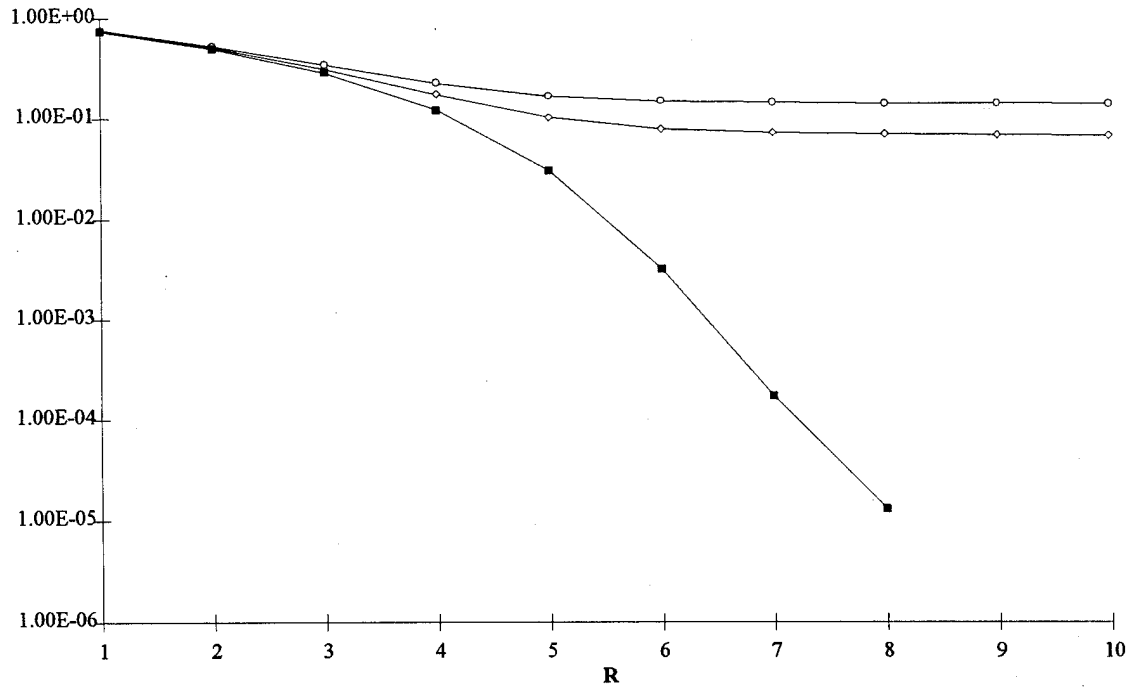


Figure 7. Cell loss rate under "hot-spot" traffic for N=1024 port tandem banyan network. The two topmost curves show cell loss rate with 5% and 10% "hot spot" traffic. The lowest curve shows uniform traffic cell loss rate and is included as a reference. Input link load is 100%.

In the next section we describe the types of synchronization operations that can be efficiently handled by the smart network. These operations are typical of distributed computing and would cause severe "hot-spot" congestion if carried over the tandem banyan network.

2.5.1. Smart Network Operations

Four types of operations are supported in the smart network architecture. These operations include one-to-one communication, broadcast communication and two types of synchronization operations. The first operation, one-to-one communication, is implemented using the tandem banyan network (see section 4) and will not be discussed here.

The broadcast operation is implemented as a user-initiated service[24]. With this approach, the originator of the broadcast transmits the master packet. This packet contains the data to be broadcast. At the same time, the recipients of the broadcast send copy packets into the network. Inside the network, the contents of the master packet are copied into the copy packets. Finally the copy packets are delivered back to their original senders. In other words, the input ports that participate in a broadcast operation, simply send packets to themselves. One packet is designated the master packet while the rest are called the copy packets. The smart network copies the contents of the master packet into the copy packets while the messages are en-route to their destination.

To allow multiple broadcast transmissions to occur simultaneously with one-to-one communication, the smart network provides special packet header fields called the group address and the COPY instruction fields. With this approach, multiple master packets can be simultaneously transmitted provided that they use unique group address values. To receive a copy of a specific master packet, the input port simply sends itself packet containing the group address of the master packet that it wishes to copy. To differentiate broadcast and one-to-one traffic, all broadcast packets contain 1 in the COPY instruction field. On the other hand, all one-to-one packets contain 0 in that field.

The remaining operations, fanin and partial-sum, are similar to broadcast transmission. Both of these operations use the group address and instruction fields in the packet header. Again, the input ports that participate in the operation, simply send packets to themselves. Their group address field is set to a unique and predetermined number while the appropriate instruction field is set to 1. The smart network then performs the requested synchronization operation while the messages are en-route to their destination.

The fanin operation allows packets that are sent to the same destination output port to be combined inside the interconnection network such that only one packet is delivered at the output. This operation is useful in distributed computing because many parallel algorithms depend on barrier synchronization[14] which require that all the processors involved in the computation send a completion status message to a specific processor to determine whether a solution has been found. If the fanin operation is not implemented in the network hardware, then packets sent to the same output port are delivered sequentially because a network output port can only accept one packet at a time. Thus when a large number of packets are sent to the same output port, a serious performance bottleneck occurs. In order to combine the packets, one needs to specify the

mathematical function that is to be performed on the packet payloads when packets are combined. Functions useful for synchronization purposes are AND, OR, MAX and MIN.

The partial sum operation allows the implementation of the fetch-and-add synchronization operation which has been found useful for many application in distributed computing[25]. The basic idea behind this operation is that the processors send a packet containing their number into the network and receive a packet that contains the partial sum of the numbers. A detailed description of the fetch-and-add operation and its usage can be found in reference 14. The absence of the partial sum operation would lead to a serious performance bottleneck, especially when large number of processors are involved.

2.5.2. Smart Network Architecture

Previously, a shuffle-based MIN architecture has been developed to support synchronization operations and to reduce internal blocking for the NYU ultracomputer project[26]. This architecture uses a bi-directional perfect shuffle MIN. Complex switching elements implement the necessary logic for performing synchronization operations and provide packet buffering in case of internal contention. Although this architecture is well suited to VLSI implementation, it is not efficient with our optoelectronic hardware module. As shown in reference 14, the use of complex switching elements in the optoelectronic MIN leads to low system performance and high hardware cost.

On the other hand, in the field of telecommunications a non-blocking interconnection network, called the STARLITE[8,10], has been developed. The STARLITE supports packet priority, one-to-one communication and broadcast communication. As shown in figure 8, the basic STARLITE design uses 4 cascaded MINs. The first three MINs (group sorting, copy and mark networks) implement broadcast communication and packet priority services. The fourth MIN is a batcher-banyan network that implements non-blocking one-to-one communication service. Although the STARLITE does not support synchronization services, it is well suited for optoelectronic implementation because it uses simple switching elements interconnected with the perfect shuffle topology.

To enable STARLITE to perform synchronization services, we add four additional MINs and replace the batcher-banyan MIN with the tandem banyan MIN. The resulting network, called the smart network, is shown in figure 8. The tandem banyan MIN portion of the smart network is called the communication section because it handles one-to-one communication traffic. The remaining group of MINs are called the processing section because they handle synchronization operations and one-to-many communication traffic.

The total number of stages in the smart network is $\log_2^2 N + (4 + R) \cdot \log_2 N + 1$ where R is the tandem banyan parameter described in section 4. For example, for a 1024 port smart network with cell loss rate of 10^{-5}, R is set to 8 and the total number of stages is 221. The worst-case latency of this network is 381 clock cycles where 240 clock cycles are spent in the tandem banyan network and the latency in the remaining networks is 1 clock cycle per stage.

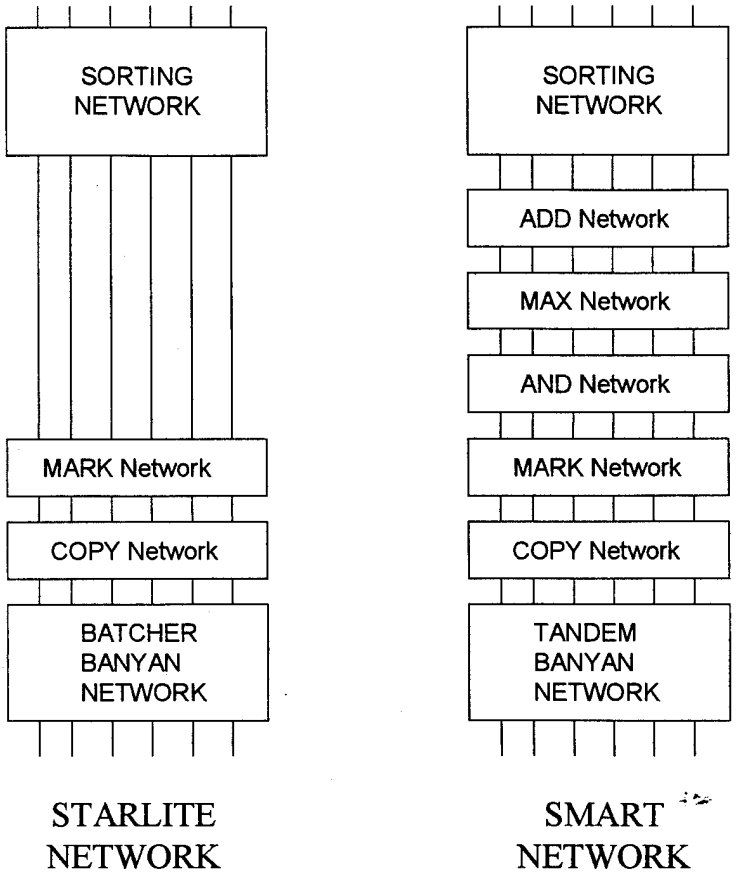


Figure 8. STARLITE and smart network architectures. The smart network adds additional stages to the STARLITE and replaces the batcher banyan with the tandem banyan.

In the smart network, packets use a complex header format with $\log_2$N bits allocated for group address, 1 bit for group priority, 6 bits for synchronization instructions, 1 bit to indicate if the packet contains active payload, 1 bit for packet priority, and $\log_2$N bits for destination address. The total size of the smart network packet header is $9+2\cdot\log_2 N$ bits. For example, a 1024 port smart network will require a 29 bit packet header. The size and format of the payload field of the smart network packet can be determined by the user. For example, to carry standard ATM payloads of 48 bytes, it should be set to 384 bits.

The group and instruction fields in the packet header specify the synchronization operation to be performed on the packets. For packets that require no internal network processing (as in the case of one-to-one communication traffic), these fields are set to 0 and the packet moves through the processing section of the smart network without being modified. Then, it enters the tandem banyan MIN where it is routed to the output port specified in the destination address field. Thus for one-to-one communication traffic, the smart network functions as a tandem banyan network with extra latency for moving packets through the processing MINs.

On the other hand, when packets do require synchronization operations, the group and instruction fields in the packet header must be set accordingly. For example, consider the case where M processors attached to the smart network request a partial sum operation. First, each processor

transmits a packet with the same predetermined number in the group address field and their own address in the destination address field. In addition, the ADD instruction field is set to 1 while other instruction fields are set to 0. The payloads contain the numbers to be added. The first MIN, the group sorting network, groups the packets based on their group address field. The second network, the ADD network, computes the partial sums of the for those packet groups that have the ADD instruction field set to 1. These networks will compute the partial sum for our M packets. Since these packets do not require other synchronization operations, they move through the remaining processing networks without being modified. Finally, the tandem banyan network delivers the M packets to their destinations (e.g. back to the senders).

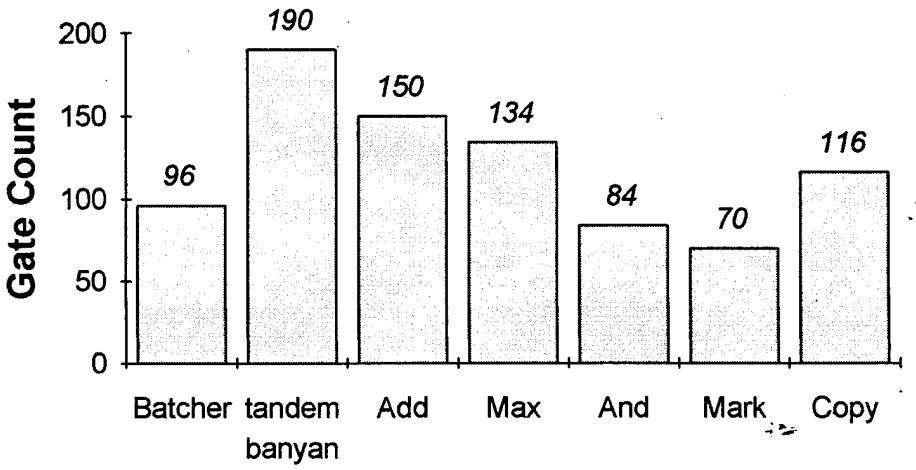


Figure 9. Design complexity for the seven types of switching elements used in the smart network. These results were obtained using VHDL synthesis.

The STARLITE architecture uses the perfect shuffle interconnection topology. The additional MINs required to implement the smart network architecture also rely on the perfect shuffle interconnection topology with the addition of some local near-neighbor interconnections in every stage. These local interconnections can be efficiently implemented using local electrical wires on the optoelectronic chips. As shown in figure 9, the 7 types of switching elements used in the smart network are simple, requiring less than 200 logic gates. Thus the smart network architecture can be efficiently implemented with our optoelectronic hardware module. The next section describes the detailed design of the smart network architecture.

2.5.3. Detailed Architecture Design

A clock-accurate, gate-level design of the smart network architecture has been developed and verified using VHDL simulation and synthesis tools. The remainder of this section describes the operation of the various networks that make up the smart network and details their design.

The first MIN in the smart network is the group sorting network[27]. This network positions packets that belong to the same group (e.g. have the same number specified in the group address field of the packet header) next to each other. This allows the processing networks that follow the sorting network to efficiently perform synchronization operations and broadcast communication. An example of a sorting network is shown in figure 10.

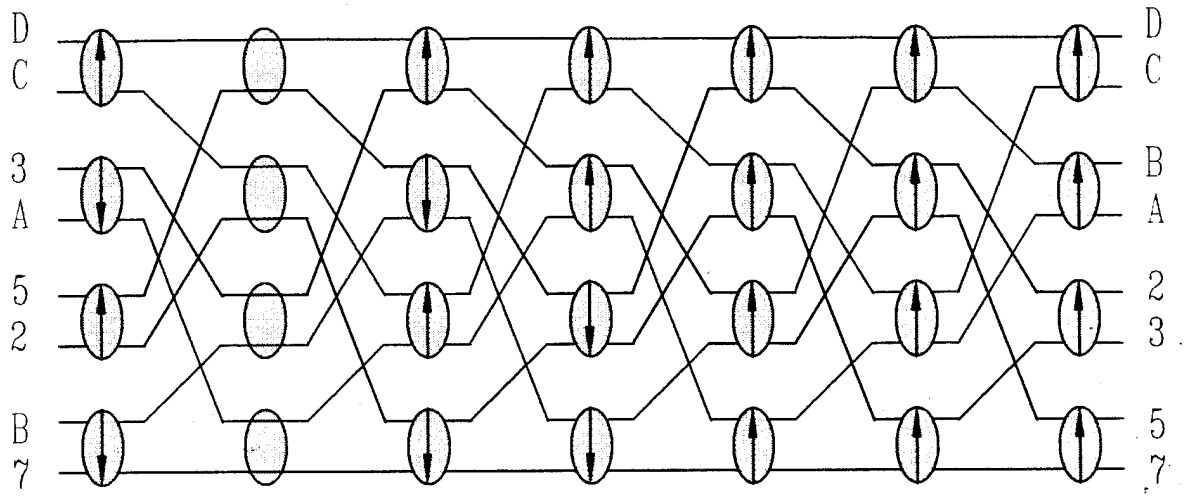


Figure 10. The sorting network groups incoming cells according to the group address field in the packet header. In this example, packets A, B, C and D have the identical group address.

The sorting network operates on the group address and group priority fields in the packet header. It is constructed using $\log_2^2 N$ stages of the perfect-shuffle interconnection network[10,28,29]. Each stage contains N/2 sorting nodes with 2 input and 2 output ports. The sorting node compares the group address fields of the two incoming cells and send the lowest numbered cell to a predetermined output port. Figure 12(a) shows the gate-level design of the switching element for the batcher sorting network. Since the design of sorting networks is widely known it will not be discussed further.

An important feature of the sorting network is that it uses both group address and group priority fields in the sorting process. Thus, within a group, packets are ordered by priority. This provides a method of controlling the order in which packets in a group are processed. For example, consider the partial sum operation (see section 5.1). Packets with higher priority will be added before packets with lower priority. Thus higher priority packets will end up having smaller partial sum values. In an application such as a parallel queue implementation[30], this mechanism can be used to implement priority services.

The group sorting network is followed by five processing networks that perform synchronization and broadcast operations. Four of these networks are constructed using $\log_2 N$ stages of the perfect-shuffle interconnection network. The fifth network, consists of a single stage that implements special internal processing. Each stage contains N processing nodes with 2 input and 1 output ports. Additionally each stage also includes local connections between the processing nodes as illustrated in figure 11. There are five different processing node types (e.g. one type for every processing network). To perform their work, processing networks must examine group address, group priority and synchronization instruction fields in the packet header. The input data and the results of synchronization operations are stored in the packet payload.

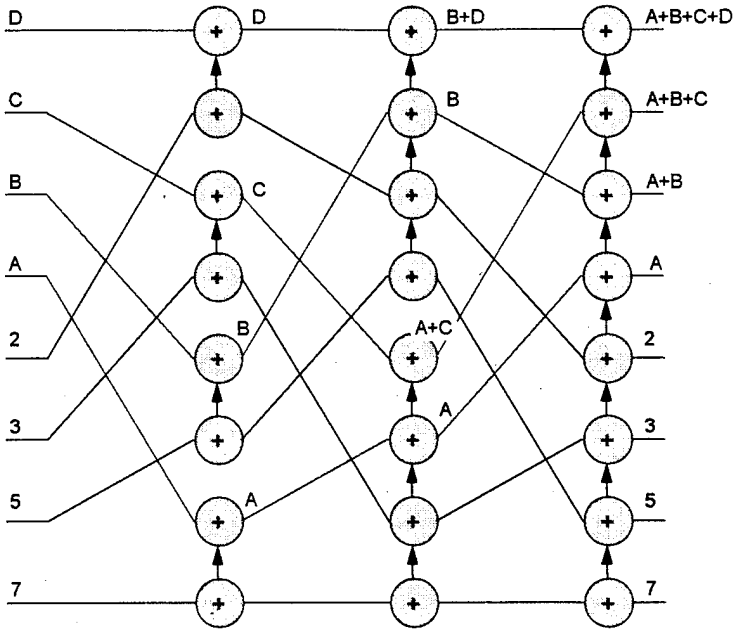


Figure 11. Interconnection network for the ADD, MAX, AND and COPY networks. This figure shows the ADD network that performs the partial add operation. Note the extra local links between switching elements.

A common feature of all the processing networks is the method used by the processing nodes to determine if the two incoming packets need to be modified. To perform this function, the processing node compares the group address fields of the two packets. If the group addresses do not match then the topmost packet is passed to the output port without modification. On the other hand, if the group addresses are identical, the processing node then checks that both packets have the appropriate synchronization instruction field set to 1. For example, the processing nodes in the COPY network check that both incoming packets have the copy instruction field set to 1. When this condition is met, the appropriate operation is performed on the payloads of the incoming packets and the topmost packet with a newly computed payload is passed to the output port. Otherwise, the topmost packet is again passed to the output port without modification.

The first processing network is the ADD network. This network implements the partial sum operation as described in section 2.5.1. Each processing node in this network includes a bit-serial adder required to compute the partial sum. At the output of this network, groups of packets that request the partial sum operation will contain the partial sum values in their payloads. The processing node for the ADD network is shown in figure 12(b).

The second processing network is the MAX network. This network implements the fanin operation based on the MAX numeric function as described in section 2.5.1. Each processing node in this network includes a bit-serial comparator. During a MAX operation, this comparator is used to compare the payloads of the two packets. Then, the largest payload value is copied to the output port. The result of the MAX operation is to copy the largest payload value within the group into the packet at the top of the group. The processing node for the MAX network is shown in figure 12(c).

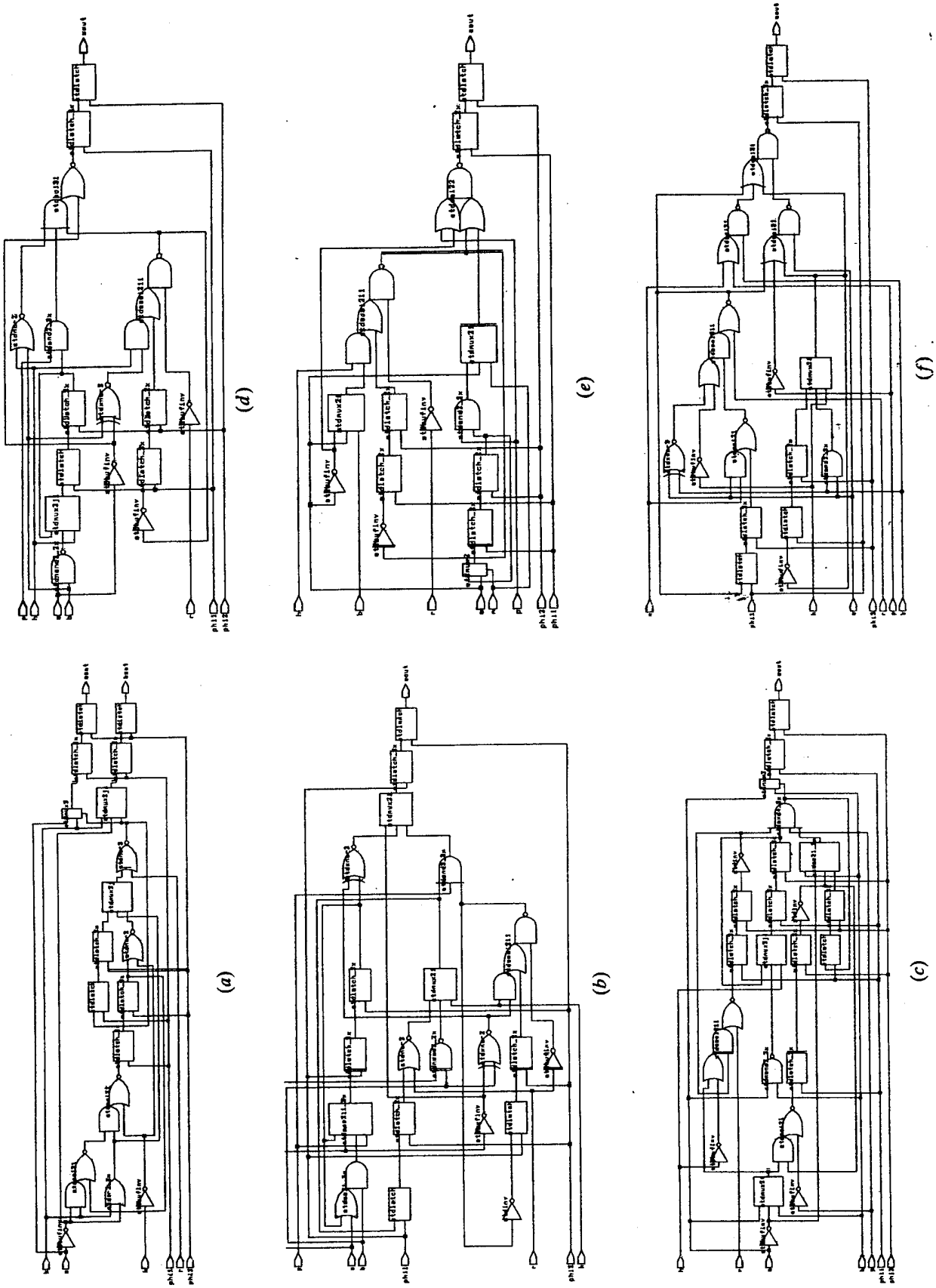


Figure 11. Gate level design for switching elements: (*a*) sorting PE; (*b*) ADD PE; (*c*) MAX PE; (*d*) AND PE; (*e*) MARK PE; (*f*) COPY PE.

The third processing network is the AND network. This network implements the fanin operation based on the AND logical function as described in section 2.5.1. Each processing node in this network includes an AND gate. During an AND operation, this gate is used to perform a bit-wise and on the payloads of the two input packets. The result is then copied to the output port. The effect of the AND operation is to copy the bit-wise and of all payloads in the group into the packet at the top of the group. The processing node for the AND network is shown in figure 12(d).

The last two processing networks are COPY and MARK networks. These two networks combine to implement the fanout operation (or broadcast communication) as described in section 2.5.1. The MARK network is used to identify and mark the master packet within a group. The COPY network is used to copy the contents of this master packet into other packets within that group. The MARK network consists of a single stage. The processing nodes for the two networks are shown in figures 12(e) and 12(f) respectively. The operation of the MARK and COPY networks has been previously described in the STARLITE design and will not be discussed further.

The last portion of the smart network is concerned with routing the cells to their final destination. This is accomplished by the familiar tandem banyan network described in section 2.4. This network uses the activity, destination address and priority fields in the packet header to route packets.

2.6. *The Hierarchical Network*

The smart network excels at performing synchronizations, however this performance comes at the expense of higher latency for one-to-one communication traffic. For example, consider a 1024 channel smart network. This network has 221 stages (100 stages for the sorting network, 41 stages for the processing networks, and 80 stages for the tandem banyan network). The lowest latency of the smart network is 171 clock cycles, compared with 30 clock cycles required in the tandem banyan network. The average latency of the smart network is 231 clock cycles versus 90 clock cycles required in the tandem banyan network.

The higher latency is acceptable in wide area networks, where the processors are separated by long distances and therefore the signal flight time delay is the dominant latency parameter. In local area networks, this latency may not be acceptable. Here we describe a hierarchical network architecture that combines the smart network with the tandem banyan network to achieve both hardware assisted synchronization operations and latency one-to-one communication. The concept of hierarchical networks has been previously proposed in reference 31.

Figure 13 shows the architecture of the hierarchical network. The basic idea is to separate incoming traffic into one-to-one communication and synchronization (including multicast) traffic. The input controller directs one-to-one communication traffic to be routed by the tandem banyan network, while the remaining traffic is directed to the smart network. With this approach the average latency for communication traffic is 90 clock cycles, while synchronization and multicast traffic have average latency of 231 clock cycles. Thus this approach offers good performance to both types of traffic present in distributed computing and can be effectively applied in the local area network environment.

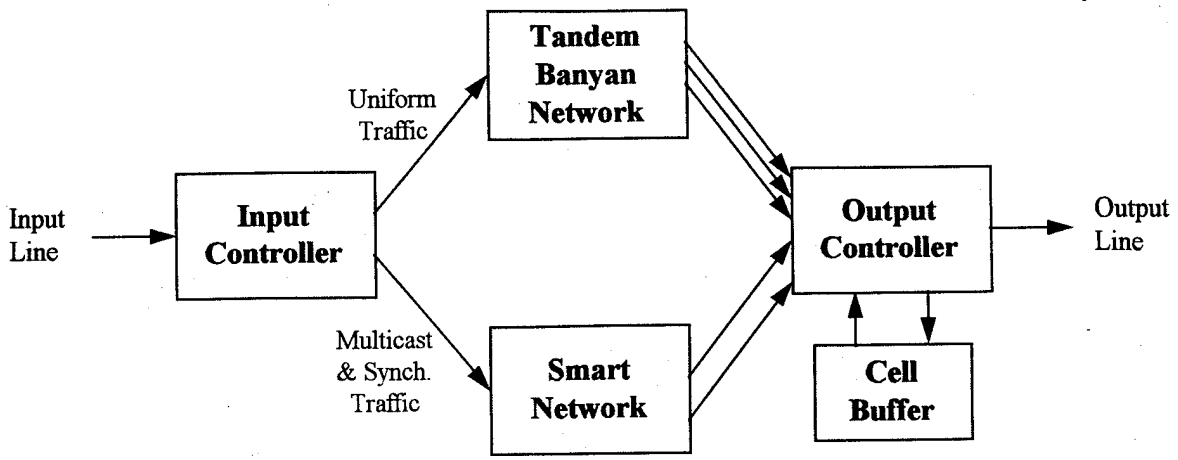


Figure 13. Hierarchical network architecture combines tandem banyan and the smart network in a parallel configuration.

2.7. *Conclusions*

In this paper, we have examined three network designs. We developed these designs for different applications and thus they have different cost, performance and functionality characteristics. A common feature of all these designs is their use of a common optoelectronic hardware module as their building block. This module implements the perfect shuffle network and has been previously shown to outperform electronic implementations at the chip and MCM levels of the packaging hierarchy[2].

	Tandem Banyan	Smart Network	Hierarchical Network
Network Size	1024	1024	1024
Cell Loss Rate for Uniform Traffic	10^{-5}	10^{-5}	10^{-5}
1-1 comm. operations	yes	yes	yes
Broadcast Operations	no	yes	yes
Synchronization Operations	no	yes	yes
Number of Stages	80	221	301
Latency (worst-case)	240	381	240 for 1-1 comm. 381 for other comm
Latency (best-case)	30	171	30 for 1-1 comm. 171 for other comm

Table 1. Summary of proposed MIN architecture designs.

Table 1 shows the number of stages, worst-case and best-case latencies for the proposed architectures when N is 1024 and cell loss rate is kept at 10^{-5}. The number of stages required for the tandem banyan network must be determined through time consuming computer simulation.

However, we expect that the relative performance and cost characteristics of these architectures will remain unmodified as N is changed.

A number of issues remain to be solved in our architecture designs. The first issue is to reduce the latency in the tandem banyan network. The second issue is to reduce "hot spot" contention that results from traffic that is not associated with synchronization operations. It turns out that both of these problems can be solved by a simple modification in the tandem banyan design. We have done preliminary work in this area and demonstrated dramatic reduction in latency at the expense of increased hardware cost[32].

Another challenge that remains to be addressed is reduction of the 3 clock cycle latency incurred in the tandem banyan switching element. In the current design, the use of longer priority fields leads to higher latency. We have also performed initial work in this area, demonstrating a design that uses parallel optical channels to transmit the packet. This design can reduce the switching element latency to one clock cycle without any appreciable increase in hardware cost[33].

It might also be argued that the optoelectronic implementation of the proposed networks would be prohibitively expensive due the large number of MIN stages and the high cost of optical interconnect devices required at every stage. However, it has been shown that a large shuffle networks can be decomposed into many smaller shuffle network interconnected with the shuffle topology[34]. In reference 14, we used this idea to partition the MIN system, implementing small electronic shuffles within a single chip and using free-space optical interconnects to link these electronic shuffles. This approach dramatically reduces the number of optical stages in the MIN. For example, the tandem banyan discussed in this paper can be implemented using 16 optical stages instead of 80 stages mentioned earlier in this paper.

The adoption of reusable generic components for building high-performance optoelectronic systems will be critical to the success of this technology. The building block approach allows the same packaging scheme and optical interconnect devices to be used in several applications, thus leveraging the development cost across a large number of potential users. In this paper we have considered an application area of switch fabrics for computer networks and shown that a generic optoelectronic hardware module can be used to implement architectures with various application specific requirements.

3. Comparison with Electronics

In this section we focus on the design of optically interconnected MCMs for gigabit ATM switching networks. Our approach is to design a generic hardware module that can be used to implement ATM switches with application-specific functionality, cost and performance requirements. The module design is partitioned on the MCM such that it can be built using VLSI chips interconnected with holographic free-space optical interconnects. Holographic optical interconnects are also used for inter-MCM communication. A comparison of our approach with electrical MCM, all-optical, and guided-wave implementations of ATM switches is presented.

3.1. INTRODUCTION

The networking industry is undergoing a dramatic change. The increasing popularity of distributed workstation computing, metacomputing, multimedia, and video teleconferencing applications coupled with the availability of 200+ MIPS workstations and gigabit fiber optic links are exhausting the capacity of present switching systems. Future network applications will require switches with large number of physical ports (1K-10K) operating at gigabit data rates (1-10 Gbps/port) and achieving terabit aggregate bandwidth capacities (1-10 Tbps).

Scaling present electronic switches to meet future networking requirements is a formidable challenge. On the technology front, physical packaging constraints of electronics (i.e. cross-talk, clock-skew, signal attenuation, limited chip pin-outs, connector limitations on PCB's, etc.) limit the connection density-bandwidth product that can be achieved within a switch (see table 1)[Error! Bookmark not defined.,35]. On the architecture front, current switch designs suffer from performance and/or cost bottlenecks if scaled beyond several hundred physical ports.

Switch	*State-of-the-art electronics*			*Optically Interconnected MCMs*		
Packaging Level	**signal density**	**data rate**	**density speed product**	**signal density**	**data rate**	**density speed product**
switch chip	64 IOs	1000 Mbps	64 Gbps/chip	1024 IOs	1000 Mbps	1024 Gbps/chip
inter-chip	800 lines/cm	500 Mbps	400 Gbps × lines/cm	10,000 lines/cm	2500 Mbps	25 Tbps × lines/cm
inter-MCM	150 lines/cm	500 Mbps	75 Gbps × lines/cm	10,000 lines/cm	2500 Mbps	25 Tbps × lines/cm
inter-board	40 lines/cm	500 Mbps	20 Gbps × lines/cm	1,000 lines/cm^2	2500 Mbps	2.5 Tbps × lines/cm^2

Table 1: Interconnect capability comparison

This paper describes how optically interconnected MCM technology can be used to build switches that will efficiently meet future networking requirements. This technology is based on combining **advanced packaging** techniques (e.g. flip-chip MCMs), **high-speed submicron VLSI** circuits (for switching) and **gigabit surface-normal optical interconnects** (for inter-chip, inter-MCM, and inter-board communication). Our approach is to develop standard optoelectronic components and packaging schemes that can be used to build high-performance application specific switches. The section is organized as follows: section 3.2 briefly reviews the technology used in our design. Sections 3.3 and 3.4 describe the organization of an ATM switching network. Section 3.5 describes previously proposed electronic ATM switch fabrics. Section 3.6 describes our design for an ATM switch fabric. Section 3.7 compares our design with electrical MCM, all-optical and guided-wave approaches. Section 3.8 provides initial results of our effort to build a prototype system based on the proposed approach. Finally, section 3.9 presents our conclusions.

3.2. OPTICALLY INTERCONNECTED MCMs

Our design uses optically interconnected multichip module technology being developed at UNC Charlotte under DARPA funding. Here we give provide only a brief description necessary to design our ATM switch system.

Figure 1 shows the proposed packaging approach, whereby multiple MCMs (called translator modules) are attached to the holographic PC board[36]. Holographic optical interconnects are used for chip-to-chip communication within and MCM as well as for MCM-to-MCM communication. Note that the same interconnect packaging is used for both levels of the packaging hierarchy thus providing similar interconnect density at both these levels. In our approach we use 2-D arrays of VSCELs directly bonded on top of the switching chips to increase the I/O capability of the switch chips.

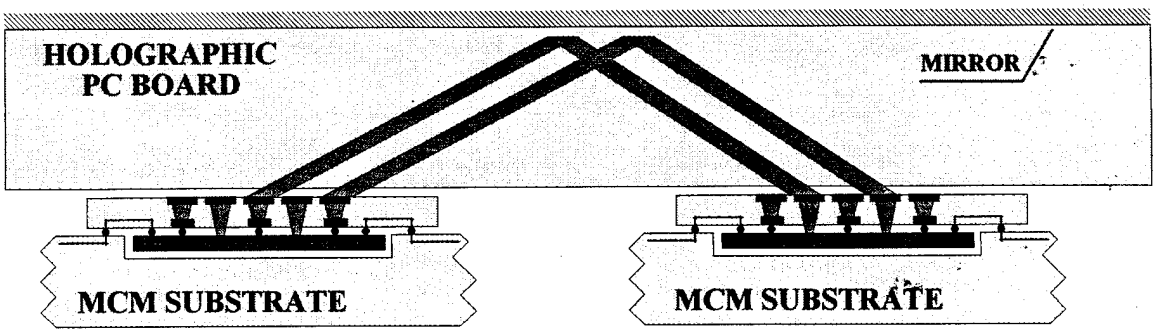


Figure 1. Multiple optically interconnected MCMs packaged on a holographic PC board.

An optical connection is made by transmitting an electrical signal from a VLSI chip through a flip-chip bonding pad and onto the translator module. The electrical signal activates a laser diode on the bottom side of the translator module. The laser generates an optical beam directed toward the holographic optical lens on the top side of the translator module. The lens collimates the optical beam and directs it to the holographic PC board. The beam is them directed onto an appropriate detector subhologram on the holographic PC board (after reflection from the planar mirror). The beam then passes through a holographic optical lens on a possibly different translator module that focuses the light onto the detector on the receiving VLSI chip. For long optical connections multiple reflections will be used to maximize the optical interconnection density. Multiple reflections can be achieved by placing metallic regions on the bottom side of the holographic PC board substrate.

3.3. NETWORK ARCHITECTURE

The high level architecture of the optoelectronic ATM network is illustrated in figure 2. It consists of multiple optoelectronic switch fabrics, input and output controllers (integrated within the buffer controller), and the system controller. Multiple switch fabrics are used for cell routing in order to improve performance and reliability. The system controller is used for higher level functions, such as network management and fabric testing. The function of the buffer controllers is to provide an external optical I/O interface, cell buffering and contention resolution mechanism.

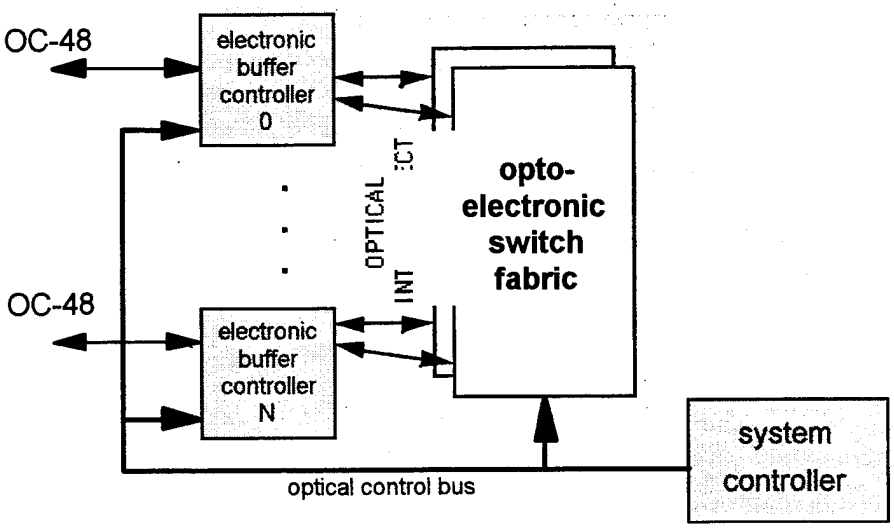


Figure 2. ATM switch architecture.

The optoelectronic switch fabrics and the buffer controllers are optically interconnected and packaged using the packaging scheme described in section 2. The use of optoelectronic packaging allows the entire switch to be packaged in a small volume, thus achieving low interconnect latencies and high clock rates. Inter-chip and inter-MCM holographic optical interconnects are used to provide the required internal wiring density and bandwidth. We note that figure 2 does not shown additional hardware modules required for interfacing OC-48 or OC-192 connection to the buffer controller. The function of these modules is to provide an ATM framer, VCI and VPI translation, and clock recovery.

3.4. SWITCH FABRIC ARCHITECTURE

A number of switch fabric architectures have been previously proposed for ATM switches. Our focus here is on the implementation of a certain class of self-routing switch fabrics, herein called banyan-based switch fabrics, that use the banyan or a functionally equivalent interconnection topology. Examples of topologies that are equivalent to the banyan include shuffle-exchange (or perfect shuffle), omega, flip, cube, and baseline[37]. Our decision to use the banyan topology is motivated by their implicit simplicity and their scalability to large number of ports.

Banyan-based switch fabrics cover a large class of switch fabric architectures, sometimes called multistage interconnection networks, that exhibit various performance and cost characteristics. The specific architecture to be used for the optoelectronic switch can be chosen to fit application-specific functionality and to optimize system performance/cost. Possible choices include cascaded banyan, tandem banyan, and batcher banyan architectures. Multiple switch fabrics can also be considered to improve network performance and reliability.

The cascaded banyan is a simple architecture requiring $N \cdot \log_2 N$ 2x2 switching elements to construct an N channel network[38]. Although it uses a distribution network to randomize the incoming cell traffic, it is an internally blocking network. Moreover, the likelihood of internal blocking increases as the switch size is scaled up. An alternative, non-blocking architecture is the batcher-banyan network[39]. The drawback of this architecture is its higher complexity, because

$N/2 \cdot (\log_2 N + 1) \cdot \log_2 N$ 2x2 switching elements are required. If some blocking is acceptable, one can design a switch fabric that uses less hardware than the batcher-banyan while achieving nearly identical performance (e.g. the amount of blocking is a design parameter). This architecture is the tandem banyan network[40].

Our design will be built using pipelined and unbuffered KxK crossbar switching elements[41,42,43], thus achieving simple and high-speed switching element design. Increasing the switching element size (K), reduces the total number of switching elements at the expense of higher individual switching element complexity. Increasing the channel width (W) allows the switch to operate at lower internal rates at the expense of higher interconnect density requirements. The switch size (K) and channel width (W) for the optoelectronic switch can be varied to optimize performance/cost for a given application.

Congestion in the switch fabric is an important design concern for gigabit switches[44]. Congestion can occur within the switch fabric due to internal contention or when several input ports send cells to the same output port. Possible solutions to congestion problem are to buffer the cells within the switch fabric or to notify the input ports of congestion. The latter approach is preferred in gigabit networks and will be used in our design.

3.5. ELECTRONIC SWITCH FABRICS

To put our approach in proper perspective, we first review electronic switch fabrics. Most electronic implementations of large switch fabrics use multiple **switch chips** interconnected on one or more PCBs or MCMs. Figure 3 illustrates this approach, showing a 64x64 (e.g. 64 channel) perfect shuffle switch fabric constructed from 16 8x8 switch chips. Each switch chip contains 12 2x2 switching elements. Each 2x2 switching element is an integrated circuit containing several hundred logic gates.

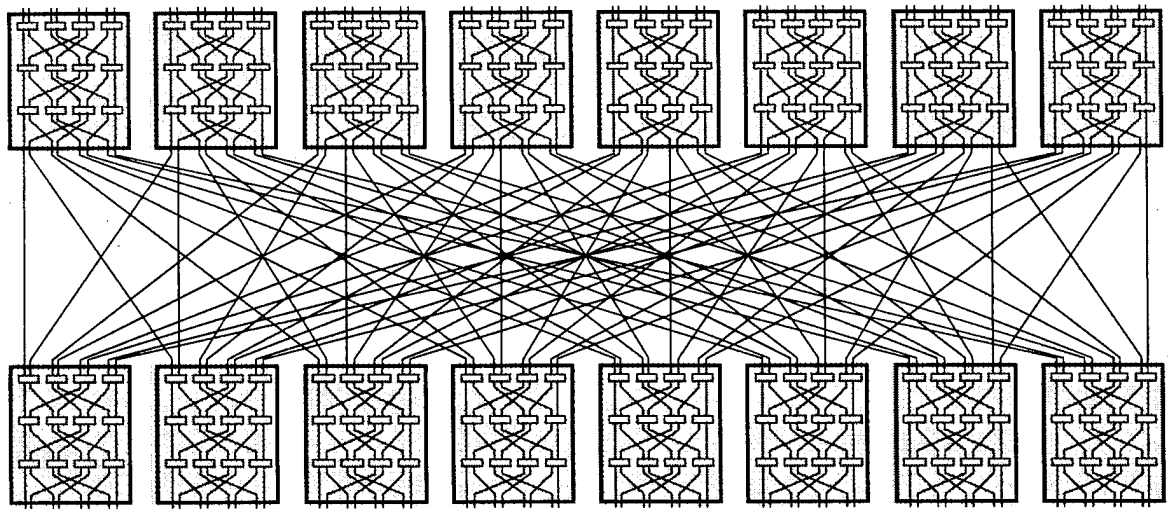


Figure 3. Planar electronic packaging scheme

A major problem with previous implementations of switch chips for banyan-based switch fabrics is the limited number of network channels that can be concurrently processed on-chip. For example, previous switch chips for batcher-banyan switch fabrics processed 32-64 on-chip network channels[45,46,47]. This limitation occurs for many reasons:

1. The use of perimeter chip-to-substrate contact technology (i.e. wire bonding, TAB) limits the contact density from the chip-to-substrate. For example, a 1cm wide chip with 100μm contact pitch has at most 400 contact pads some of which have to be used for power and control signals. This leaves less than 400 contact pads for network channels. We note that although array chip-to-substrate contact technology (i.e. flip-chip) can be used here to increase the chip pin-outs, taking advantage of these additional pins will be difficult in light of chip power consumption and substrate connection density constraints.

2. In high-speed chip designs, the chip power consumption is dominated by the output drivers that are required to drive transmission lines. The power required to drive a transmission line is about 10mW (e.g. $V^2/R=1^2/100$). Assuming a 5 watt power budget for the output drivers, at most 500 output channels are possible.

3. Previous chip layouts of banyan and its equivalents used one-dimensional layouts[48], that do not scale well in area and wire length when the number of on-chip channels is large. With 1μm design rules, chip area and chip speed become wire-limited when the number of on-chip channels is increased beyond 256.

The limited number of on-chip network channels implies that a large number of switch chips will be required to implement a large switch fabric. For example, over 700 chips will be required to implement a 1024x1024 batcher-banyan switch fabric using 32x32 switch chips (i.e. $\left[\#\ stages\right] \times chips/stage = \left[\log_{32} 1024 \cdot \left(\log_2 1024 + 1\right)\right] \times 1024/32 = 704$). Even with 64x64 switch chips, the total number of chips will still be about 300. This situation creates the following problems:

1. A system with a large number of chips has high fabrication cost, low reliability, long off-chip wire lengths (and hence reduced clock speed), high power consumption, and large system size. If multiple PCBs or MCMs are required to contain all the chips, the situation becomes even worse because of the higher cost and lower density that is associated with using another level of the packaging hierarchy (i.e. inter-PCB or inter-MCM).

2. Present schemes for interconnecting switch chips on the PCB or MCM require a large number of wires to cross over each other and to extend across the entire board. The relatively large interconnect pitch of PCB and MCM (typically 25μm for MCM-D and 250μm for PCB vs. 2μm on-chip) and the limited number of wiring layers (typically 2 for MCM-D and 8 for PCB) can create a wiring congestion when interconnecting a large number of switch chips.

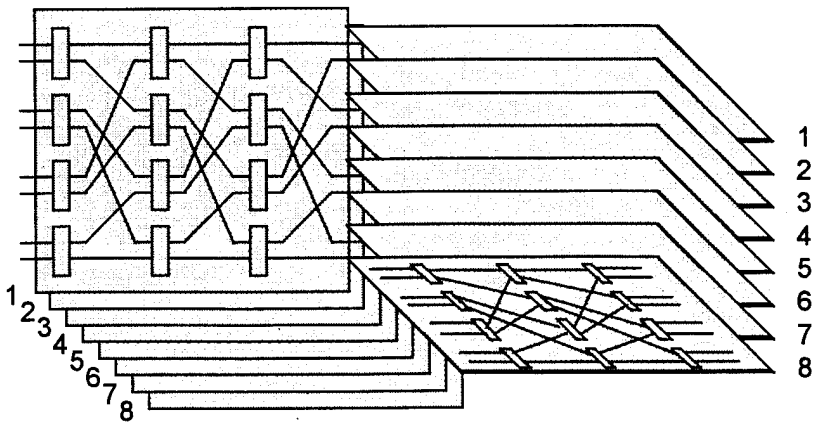


Figure 4. 3-D electronic packaging scheme

A **three-dimensional packaging scheme** for implementing banyan-based switch fabrics has been previously proposed to overcome the problems of large system size, wiring congestion and long off-chip wire lengths (see figure 4)[49]. This packaging scheme works by arranging boards with multiple switch chips into columns and interconnecting them with orthogonal planes. The main disadvantage of the 3-D electronic packaging scheme is the large number of switch chips required. This leads to high fabrication costs, and high power consumption, and reduced reliability. The next section describes the proposed approach based on 3-D optoelectronic packaging. In this scheme, the number of switch chips required is dramatically reduced over 3-D electronic packaging approach, because more network channels are processed on-chip.

3.6. Optoelectronic switch fabric

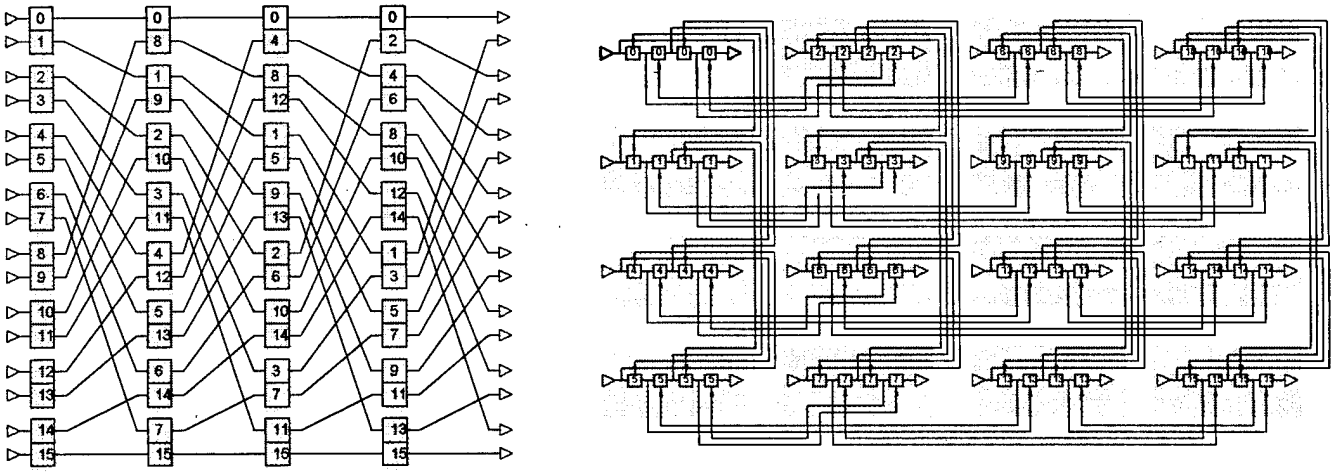
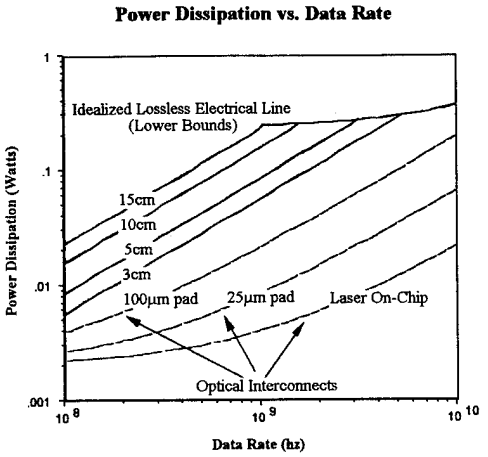
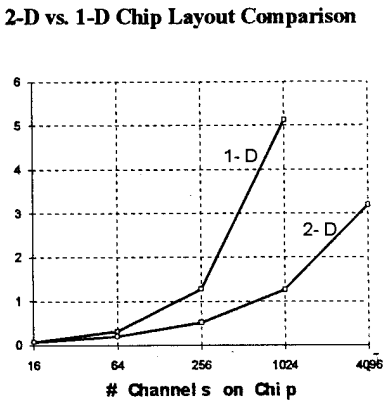


Figure 5. 2-D optoelectronic chip layout scheme

The previous section has identified the limited number of network channels in previous switch chip designs as the main impediment to efficient implementation of large banyan-based switch fabrics. In order to increase the number of on-chip network channels we propose to use a two-dimensional chip layout scheme, herein called **2-D banyan layout**. As shown in figure 5 this layout scheme is functionally equivalent to previous 1-D banyan layouts[50]. However, by uniformly distributing the I/O ports throughout the chip area, the 2-D layout scheme can accommodate

larger and faster switches within a given area. The use of optical interconnect allows more pinout to handle the increased I/O requirements. These last two points are illustrated graphically in figure 6. We note that this section assumes bit-serial channels (W=1) and 2x2 switching elements (K=2) to simplify the discussion.





- **Electrical Pin-Out**
  - **1000 IOs = 50-250 Watts**
- **Optoelectronic Pin-out**
  - **1000 IOs = 4-20 Watts**

Process technology	1.2μm CMOS Double-level metal
Minimum feature size (λ)	0.6μm
Transistors/switching element	100
Area/transistor (A_t)	555 (20μm x 10μm)
Switching element pitch (P)	0.5
Wire pitch (W)	4 (2.4μm)

Figure 6. Comparison of 2-D optoelectronic and 1-D electronic switch chip layouts

A large switch fabric requires multiple banyan networks to be interconnected in series and/or in parallel. We can achieve this by optically interconnecting multiple 2-D layout switch chips packaged on single or multiple MCMs as shown in figure 7. In this approach, several banyan networks are packaged on a single holographic PC board. The 2-D layout scheme can be repeated at higher levels of the packaging hierarchy to build larger networks. Alternatively, series interconnection of switch chips or MCMs can be used to achieve many multiple stages required for tandem-banyan and batcher-banyan networks. Inter-chip communication within the MCM can be optical or electrical depending on interconnect speed and density consideration. Our approach will be to use electrical interconnects for short lines that can be treated as lumped capacitors. Multiple PC boards are stacked in parallel to construct the complete switch fabric. Inter-board communication is done with surface normal optical interconnects. The specific partitioning of the switch architecture onto the our packaging scheme is determined by the application and the specific performance/cost requirements.

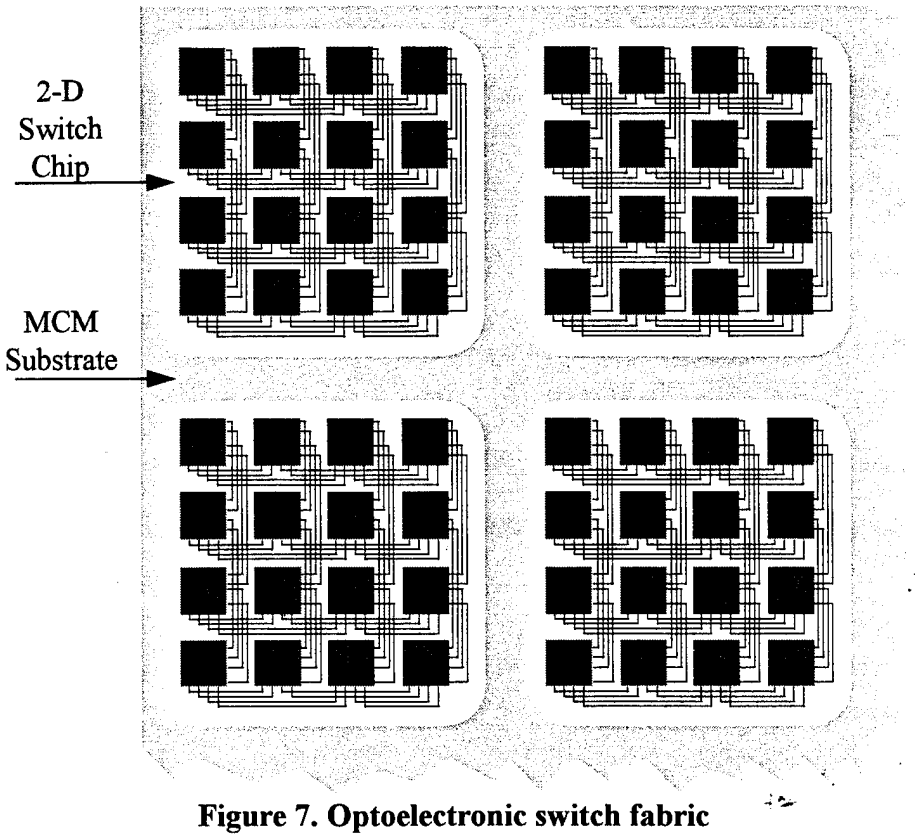


Figure 7. Optoelectronic switch fabric

Our packaging scheme provides the following advantages:

1. The 2-D layout is ideally suited to array contact pad capability of MCMs and 2-D optical I/O because it uniformly distributes the input and output ports throughout the chip area. With array contacts, the I/O pad density limit of present designs is removed. For example, a 1cm diameter chip with 100m contact pitch has at most 10,000 contact pads vs. 400 for the perimeter contact case.

2. The use of 2-D layout allows larger networks to be concurrently processed on-chip. Moreover, these networks can operate at higher speed because of shorter on-chip wires. For example, for a 1024 channel switch chip in 1μm design rules, the 2-D layout chip is a square (1.25cm x 1.25cm) with 0.63 cm longest wire length. On the other hand, the 1-D layout chip is a rectangle with impossible-to-fabricate dimensions (5.1cm x 0.45cm) with 2.70 cm longest wire length.

3. The number of chips required to implement a large network is dramatically reduced. For example, a 1024 channel batcher-banyan network requires only 11 1024 channel switch chips vs. 300 chips required with previous designs using 64 channel switch chips. In this case, the entire network can be easily achieved on a single holographic PC board. This reduces the system size, power consumption and cost.

4. The use of 3-D packaging allows dramatic reduction is system size and interconnect latencies involved in packaging many chips required for a large switch.

5. Electrical interconnects are used for on-chip wiring and for short inter-chip connections on the MCM. The use of electrical interconnects at these packaging levels is advantageous because high interconnect density and low power consumption can be achieved with electrical interconnect for on-chip wiring as well as for short inter-chip connections (e.g. short inter-chip connections do not behave as transmission lines and thus do not have the power consumption and fabrication complexity).

6. Optical interconnects are used for long chip-to-chip connections on the MCM, inter-MCM connections, and inter-board connections. The use of optical interconnect at these packaging levels allows higher connection density and lower power consumption than that possible with electrical interconnects as described in section 3.

7. The proposed scheme has excellent scalability potential. For example, with 1μm design rules, we can implement a 4096 channel cascaded banyan network using 48 1024 channel switch chips on a single holographic PC board. Multiple 4096 channel cascaded banyan network can be achieved by stacking holographic PC boards with the 3-D packaging scheme.

3.7. COMPARISON WITH OTHER APPROACHES

3.7.1. Electronic MCMs

To determine the usefulness of our approach we have compared it with an equivalent electronic implementation using state-of-the-art flip-chip electronic MCMs. Table 2 shows the assumptions made for electronic MCMs. Our electronic design uses 2-D layout switch chips interconnected using 2-D layout on the MCM (e.g. replacing optical interconnects in section 6 with electrical wires). Figure 8 shows the cost of implementing a 4096 channel banyan switch using electronics.

TECHNOLOGY:	SYSTEM DESIGN:
THIN FILM HYBRID MCM TECHNOLOGY	SHUFFLE NETWORK WITH 4096 CHANNELS, 12 STAGES
CMOS 1.2u CHIP TECHNOLOGY	REQUIRES TOTAL OF 64 CHIPS
ON-CHIP WIRING PITCH = 4 microns	EACH CHIP IS A 64 NODE HYPERCUBE
OFF-CHIP WIRING PITCH = 50 microns	768 I/O PINS PER CHIP
2 LAYERS OF SIGNAL INTERCONNECT	1.2K GATES PER NODE
CMOS OFF-CHIP DRIVERS HAVE 3V SWING	77K LOGIC GATES PER CHIP
FLIP-CHIP MOUNTING WITH 100um PAD PITCH	5M LOGIC GATES IN SYSTEM
	25K SIGNAL WIRES ON MCM
	MCM BISECTION WIDTH = 4096 WIRES

Table 2. Electronic MCM assumptions

Our results show that the electronic system power budget is dominated by the power required to drive the chip-to-chip electrical wires which behave as transmission lines. This system power bottleneck is greatly eased by the use of optical interconnects which as shown in figure 6 consume much lower power than their electrical counterparts. Likewise, the system size is dominated by the MCM substrate wiring (e.g. chips have to be widely space apart to accommodate the necessary chip-to-chip wiring). In this case, the higher density of optical interconnects allows a

more compact package to be implemented than possible with electronics. Finally, the system clock budget is dominated by the driver latency which can be expected to reduce with lower power optical interconnects. Our comparison is done at the MCM level, but as shown in table 1, we expect that the benefits of using optical interconnects will be even greater for higher levels of the packaging hierarchy.

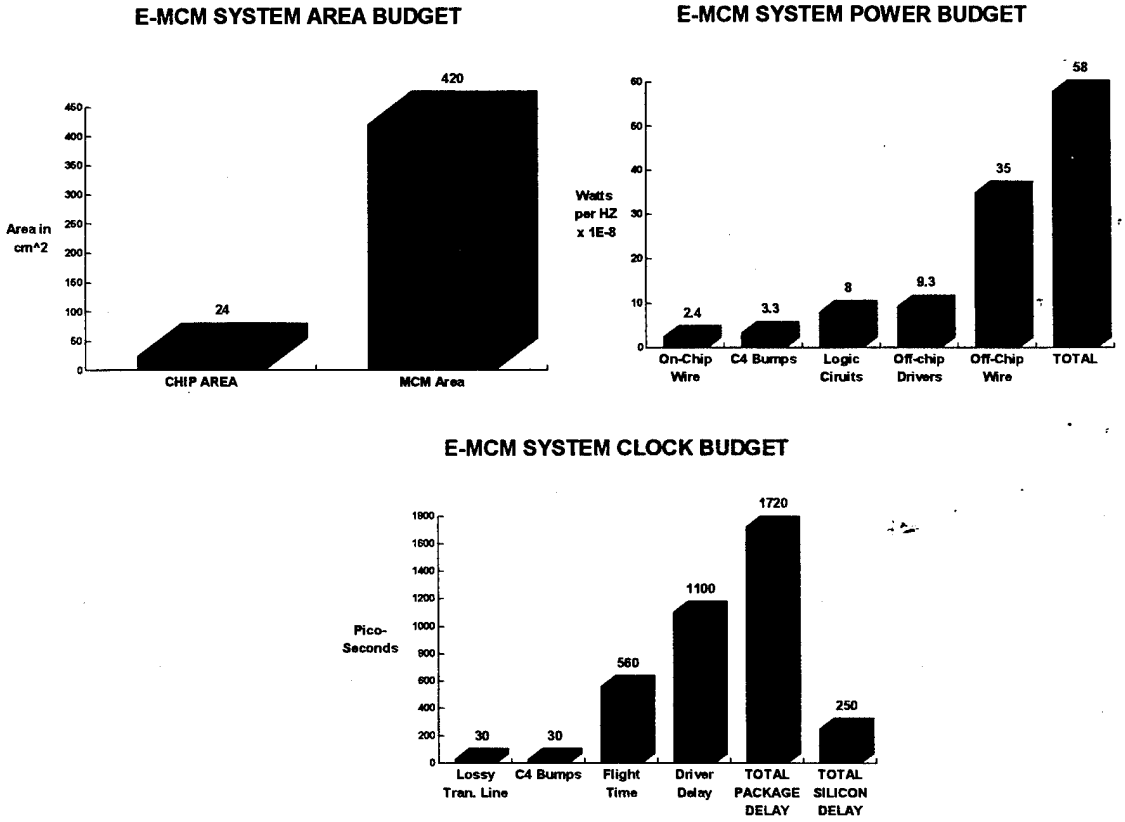


Figure 8. Comparison summary

3.8. CONCLUSION

This section has presented generic hardware module for building gigabit ATM switches. The design is based on optically interconnected MCM technology that proivdes the dense and high-speed I/O capability at multiple levels of the packaging hierarchy to meet the demands of the ATM switching application. We have compared our approach with electronic MCM, waveguide and all-optical switches showing that for large number of channels operating at gigabit channel data rates, our approach becomes attractive from performance and cost considerations.

3.9. References

1. F. Kiamilev, P. Marchand, A.V. Krishnamoorthy, S. Esener and S.H. Lee, Performance comparison between optoelectronic and VLSI multistage interconnection networks, IEEE Journal of Lightwave Technology, Vol. 9, No. 12, pp. 1674-1692, Dec. 1991.

2. F. Kiamilev, et.al., "Optically interconnected MCMs for gigabit ATM switches," in Proc. SPIE Conf. 1849 on Optoelectronic Interconnects (OE/LASE'93), paper 1849-23, (1993).

3. A.L. Decegama, *Parallel processing architectures and VLSI hardware,* (Prentice-Hall, 1989).

4. H. J. Siegel, *Interconnection networks for large-scale parallel processing*, 2nd Edition, (Mc-Graw Hill, 1990).

5. R. Nordin, A. Levi, R. Nottenburg, J. O'Gorman, and R. Logan, "A systems perspective on digital interconnection technology," IEEE J. of Lightwave Technology, Vol. 10, No. 6, pp. 811-827 (1992).

6. J.E. Midwinter and M.G. Taylor, "The reality of digital optical computing," IEEE LCS Magazine, pp. 40-46, (May 1990).

7. S.H. Lin, T.F. Krile, and J.F. Walkup, "2-D optical multistage interconnection networks," in Proc. SPIE, Digital Optical Computing, Vol. 752, pp. 209-216, (1987).

8. C.W. Stirk, R.A. Athale, and M.W. Haney, "Folded perfect shuffle optical processor," Appl. Opt., Vol. 27, No. 2, pp. 84-85, (1988).

9. A.A. Sawchuck and I. Glaser, "Geometries for optical implementations of the perfect shuffle," in Proc. SPIE, Optical Computing, Vol. 963, pp. 270-279, (1988).

10. K.H. Brenner and A. Huang, "Optical implementations of the perfect shuffle interconnection," Appl. Opt., Vol. 27, pp. 135-137, (1988).

11. S. Bian, K. Xu, and J. Hong, "Optical perfect shuffle using wollaston prisms," Appl. Opt., Vol. 30, pp. 173-174, (1991).

12. J.E. Midwinter, "Photonics in switching: the next 25 years of optical communications," IEE Proc., Vol.. 139, No. 1, pp. 1-12 (1992).

13. F.B. McCormick, et.al., "Six-stage digital free-space optical switching network using symmetric self-electro-optic-effect devices," Appl. Opt., Vol. 32, No. 26 (1993).

14. A. Krishnamoorthy, P. Marchand, F. Kiamilev, and S. Esener, "Grain-size considerations for optoelectronic multistage interconnection networks", Appl. Opt., Vol. 31 #26, 5480-5507 (1992).

15. Martin De Prycker, "Asynchronous Transfer Mode," Ellis Horwood Limited (1993).

16. N. Yamanaka, S. Kikuchi, and T. Takada, "A 1.8-Gb/s GaAs optoelectronic universal switch LSI with monolithically integrated photodetector and laser driver," IEEE J. of Lightwave Tech., Vol. 8, No. 8, pp. 1162-1166 (1992).

17. Y. Iseki, F. Shimizu, and T. Sudo, "Multichip module technology using AIN substrate for 2-Gbit/s high-speed switching module," in Proc. 42th Electronic Components and Technology Conf., pp. 973-978 (1992).

18. Y. Doi, H. Yamada, S. Sasaki, "An ATM switch hardware technology using multichip packaging," in Proc. 42th Electronic Components and Technology Conf., pp. 984-990 (1992).

19. M. Dubois, C. Scheurich, and F. Briggs, Synchronization, coherence, and event ordering in multiprocessors, IEEE Computer 21 (February 1988).

20. G.F. Pfister and V.A. Norton, "Hot spot contention and combining in multistage interconnection networks," IEEE Trans. on Computers C-34, 943-948 (October 1985).

21. H.S. Stone, High performance computer architecture, (Addison-Wesley 1980).

22. C.P. Kruskal, L. Rudolph, and M. Snir, "Efficient synchronization on multiprocessors with shared memory," ACM Trans. on Prog. Lang. and Systems 10, 579-601 (October 1988).

23. F.A. Tobagi, T. Kwok, F.M. Chiussi, "Architecture, performance, and implementation of the tandem banyan fast packet switch," IEEE J. on Sel. Areas in Communications 9, 1173-1193 (1991).

24. A. Huang, "The relationship between STARLITE, a wideband digital switch, and optics," in Proc. ICC'86, Toronto, Canada, 1725-1729 (June 1986).

25. A. Gottlieb, B.D. Lubachevsky, and L. Rudolph, "Basic techniques for the efficient coordination of very large numbers of cooperating sequential processors," ACM Trans. on Prog. Lang. and Systems 5, 164-189 (April 1983).

26. A. Gottlieb, R. Grishman, C. Cryskal, K. McAuliffe, L. Rudolph, and M. Snir, "The NYU ultracomputer - designing an MIMD shared memory parallel computer", IEEE Trans. on computers C-32, 75-89, (February 1983).

27. K.E. Batcher, "Sorting networks and their applications," 1968 Spring Joint Computer Conf., AFIPS Proc. 32, Washington D.C., 307-314 (1968).

28. H.S. Stone, "Parallel processing with the perfect shuffle," IEEE Trans. on Computers C-20, 153-161, (February 1971).

29. J.N. Giacopelli, J.J. Hickey, W.S. Marcus, W.D. Sincoskie, and M. Littlewood, "Sunshine: a high-performance self-routing broadband packet switch architecture," IEEE J. on Sel. Areas in Communications 9, 1289-1298 (1991).

30. H.S. Stone, "Database applications of the fetch-and-add instruction," IEEE Trans. on Comput C-33, 604-612 (1984).

31. P.A. Francaszek, "Path hierarchies in interconnection networks," IBM J. Res. Develop. **31**, 120-131 (January 1987).

32. A.V. Krishnamoorthy, F. Kiamilev, and S.C. Esener, "A class of packet-switched extended generalized shuffle networks," OSA Annual Meeting Technical Digest, (Optical Society of America, Washington D.C. 1992), Vol. 18, pg. 192.

33. F. Kiamilev, C. Graham, and E. Stevens, "Optoelectronic Multislotted Interconnection Networks," OSA Annual Meeting Technical Digest, (Optical Society of America, Toronto, Canada 1993), Vol. 16, pg. 76.

34. S.C. Knauer, J.H. O'Neill, and A. Huang, "Self-routing switching network," in Principles of CMOS VLSI Design, N. Weste and K. Eshraghian, ed., (Addison-Wesley 1988).

35. R. Nordin, A. Levi, R. Nottenburg, J. O'Gorman, and R. Logan, "A systems perspective on digital interconnection technology," IEEE J. of Lightwave Technology, Vol. 10, No. 6, pp. 811-827 (1992).

36. C. Osbourn, A. Reisman, L.T. Hwang, "Apparatus for mounting a semiconductor chip and making electrical connections thereto," U.S. Patent #4,744,630, (1988).

37. H. Ahmadi and W. Denzel, "A survey of modern high-performance switching techniques," IEEE Trans. on Sel. Areas in Comm., Vol. 7, No. 7, pp. 1091-1103 (1989).

38. F. T. Leighton, Introduction to parallel algorithms and architectures, Morgan-Kaufman Publishers (1992).

39. A. Huang and S. Knauer, "STARLITE: A wideband digital switch," in Proc. Globecom'84, Atlanta, GA, 121-125 (1984).

40. F.A. Tobagi, T. Kwok, F.M. Chiussi, "Architecture, performance, and implementation of the tandem banyan fast packet switch," IEEE J. on Sel. Areas in Communications 9, 1173-1193 (1991).

41. D.M. Dias and J.R. Jump, "Analysis and simulation of buffered delta networks," IEEE Trans. Comput C-30, 273-282 (1981).

42. C.P. Kruskal and M. Snir, "The performance of multistage interconnection networks for multiprocessors," IEEE Trans. Comput. C-32, 1091-1098 (1983).

43. S.C. Knauer, J.H. O'Neill, and A. Huang, "Self-routing switching network," in Principles of CMOS VLSI Design, N. Weste and K. Eshraghian, ed., Addison-Wesley (1988).

44. H.T. Kung, "Gigabit local area networks: a systems perspective," IEEE Comm. Mag., pp. 79-89 (April 1992).

45. J. Berthold, "Broadband electronic switching," IEEE LCS Mag., pp. 35-39 (May 1990).

46. W. Marcus, "A CMOS batcher and banyan chipset for B-ISDN packet switching," IEEE J. of Solid-State Circuits, Vol. 25, No. 6, pp. 1426-1439 (1990).

47. D. Boyer and R. Cordell, "Rapid prototyping of high-speed communications chips," IEEE Design and Test of Comp. 8, 27-39 (1991).

48. F. Kiamilev, P. Marchand, A.V. Krishnamoorthy, S. Esener, and S.H. Lee, "Performance comparison between optoelectronic and VLSI multistage interconnection networks", IEEE Journal of Lightwave Technology, Vol. 9, No. 12, pp. 1674-1692, Dec. 1991.

49. J. Hui, "Switching integrated broadband services by sort-banyan networks," Proc. of the IEEE, Vol. 79, No. 2 (1991).

50. A. Krishnamoorthy, P. Marchand, F. Kiamilev, and S. Esener, "Grain-size considerations for optoelectronic multistage interconnection networks", Appl. Opt., Vol. 31 #26, 5480-5507 (1992).